229
27.8.7

THE
BRITISH JOURNAL
OF
EDUCATIONAL PSYCHOLOGY

Edited by

L. B. BIRCH

Associate Editors:

P. E. VERNON and DORIS M. LEE

Advised by

SIR CYRIL BURT
R. A. C. OLIVER
E. A. PEEL
SIR FRED J. SCHONELL
C. W. VALENTINE
S. WISEMAN

Vol. XXXIII

1963

BUREAU OF EDNL. & PSYCL. RESEARCH.
LIBRARY.

Published by
METHUEN AND CO., LTD., 36, ESSEX STREET, STRAND, LONDON, W.C.2.,
for the British Journal of Educational Psychology, Ltd.

Registered Offices: 42, Hill Street, Birmingham, 5. This address should *not* be used for letters or books intended for the Editor, Authors, or the Publishers.

On behalf of

The British Psychological Society and the Association of Teachers in Colleges and Departments of Education.

All rights reserved.

PRINTED BY THE BIRMINGHAM PRINTERS, LTD., 42, HILL STREET, BIRMINGHAM, 5.

Bureau Edni. Research
DAV [illegible] COLLEGE
Dated 27. 8. 70
Acces No. J 225

EDITORIAL

THESIS ABSTRACTS

Since 1938, this *Journal* has regularly published abstracts of appropriate theses which have been accepted for higher qualifications in psychology and education. In recent years, however, in spite of the increased output of theses, these have come mainly from a very small number of universities. This note is included to remind Heads of Departments of Psychology and of Institutes and Departments of Education of the service which is available.

Thesis abstracts for consideration should be forwarded to the Editor by the Head of the Department concerned and may be intended either as the sole published account of the thesis or as a preliminary report of a research which is intended to be published at length at some future date. It is essential that they should be abstracts, not short articles, and Heads of Departments are specially asked to sub-edit them to make them as clear as possible consistent with brevity. Normally, abstracts of more than 1,500 words will not be considered and only exceptionally will tables be allowed.

Proofs will normally be returned to the Department of origin for correction.

CORRECTION

The heading of the three papers on researches using the Maudsley Personality Inventories in the last issue of the *Journal* (p. 241) should have been " Three Experiments using Maudsley Personality Inventories " not " Children's Personality Inventories." Savage used the M.P.I. in his research.

Dr. Warburton has also pointed out that the statement by Callard and Goodfellow that the J.M.P.I. is the only test of this kind available for English children, is not strictly true. Cattell's High School Personality Questionnaire (H.S.P.Q.) for ages 12-16 and his Children's Personality Questionnaire (C.P.Q.) for ages 8-12, are available in this country. Although they were standardised in the United States they have been used in Manchester and have been found suitable for use with English children.

VOLUME XXXIII FEBRUARY, 1963 PART I

A FURTHER STUDY OF THE EDUCATIONAL PROGRESS OF CHILDREN WHO HAD RECEIVED REMEDIAL EDUCATION

By K. LOVELL, C. BYRNE and B. RICHARDSON
(*University of Leeds*)

SUMMARY. Children who had received full-time teaching in Remedial Centres of an English education authority were traced, and their reading ages assessed along with their understanding of certain arithmetic concepts. A Summary of the data obtained is given together with the inter-relationships found between a number of the relevant variables.

I.—INTRODUCTION.

IN an earlier paper (Lovell, *et al.* 1962) some details were given relating to the reading ages of children who had received remedial teaching and were subsequently followed-up. The pupils concerned in that study had received one period of remedial teaching per week, each period lasting 45 minutes. In the present study junior school children in a different education authority, who had been sent to full-time remedial centres for a period, were followed-up and their reading ages assessed. Furthermore, these children also undertook a test of arithmetic concepts at follow-up.

It must be clearly stated at the outset that this investigation throws no light on whether the children who attended these centres became better adjusted to school work and life, or whether their general motivation and morale improved under full-time remedial teaching. But the study does throw light on the likely educational progress and eventual performance, during school years, of children considered to be in need of remedial teaching. Furthermore, it is now possible to compare the progress of these children with those in the earlier study who had remedial teaching for only one period a week, and it has been possible to throw some further light on the question as to whether, in the long run, children who receive remedial teaching in special centres make more progress than those who stay in their own schools.

II.—SAMPLE STUDIED AND PROCEDURE.

The local education authority had five full-time remedial centres and the teachers-in-charge visited the local junior schools and picked out children, using intelligence and reading tests, who were considered to be in need of remedial teaching. Not all children so picked out attended the centres. Sometimes, the head of a junior school did not regard a particular child as suitable, and some parents refused permission for their children to attend the centres in lieu of school. In general, those who attended the centres were deemed to be the most likely to make most improvement under special teaching.

An examination of the records at the centres yielded complete data on 261 children who had attended full-time, and who had returned to ordinary schools not later than July, 1960. The follow-up study was done during the period November, 1961—March, 1962, so that no child was included in the survey unless he had been back at ordinary school for at least sixteen months.

The subjects were then divided up into those who were backward and those who were not, the former group being defined as those pupils whose reading ages (on whatever test was used at the time of referral) were less than 85 per cent. of their chronological ages. There was thus found to be 184 children classified as backward, and 77 as not backward. It will be quite obvious at this point that some children had been selected for attendance at these centres on the grounds of a discrepancy between reading age and mental age. Of the backward children, nine had moved to other parts of the country, one was in hospital, four were absent from school during the follow-up period, and six could not be traced by the authority, giving 164 backward children available for study. Of the non-backward group only one child was not found (she had moved to a private school outside the authority's area) giving 76 pupils in this category available for investigation.

The teachers-in-charge of the centres were free to use intelligence or reading tests (often the N.F.E.R. Sentence Reading Test) of their own choice in selecting children for the centres. Of the backward children fifty-nine had taken the Moray House Picture Test 1, fifty-four the Intelligence Tests for young children (Valentine), twenty-five the Sleight Non-Verbal Intelligence Test, fourteen the N.F.E.R. Non-Verbal Test 5, and twelve the Terman-Merrill revision of the Binet test. Again, either just before or immediately after admission, 142 subjects had taken the Schonell Graded Vocabulary Test R1 and twenty-two the Holborn Reading Scale for Infants and Juniors ; while at the time of return to school, 102 had been given the Schonell Graded Vocabulary Test R1, and sixty-two the Holborn Reading Scale.

In the non-backward group forty-eight had taken the Moray House Picture Test 1, eleven the Intelligence Tests for Young Children (Valentine), three the Sleight Non-Verbal test, thirteen the N.F.E.R. Non-Verbal Test 5, and one the Terman-Merrill revision of the Binet test. Just before or immediately after admission to the centres, sixty-two children had their reading ages assessed using the Schonell Graded Vocabulary Test R1 and fourteen the Holborn Reading Scale, while on return to school the numbers for these two tests were respectively fifty-seven and nineteen.

On follow-up, the children were found in over fifty schools. Both the Schonell Graded Vocabulary Test R1, and the Vernon Graded Word Reading Test were given to assess the reading ages. Moreover, each pupil undertook, individually, a test which involved the understanding of certain arithmetic concepts devised by one of the authors (K.L.). The questions, which were duplicated were presented to the subjects. The tester read aloud a question, and the child followed on the paper. The latter could either answer orally or in writing. No time limit was imposed. When the tester was satisfied that the child had answered the question to the best of his ability the experimenter passed on to the next question. The percentage numbers of children responding correctly to some of the least well answered questions are shown in Table 4.

In order to have a further check on educational progress, figures will be presented showing the numbers of children in the bottom stream, next from bottom stream, etc., of multi-stream schools, on follow-up (Table 5).

At one centre initial data were obtained on fifty children who, on the basis of a discrepancy between R.A. and M.A., had been regarded as being suitable for attendance at the centre, but who, for the kind of reasons already explained, did not, in fact, attend. Such children were left in ordinary schools and will be referred to as the control group. All children who were picked as likely to be suitable for this particular centre had taken the N.F.E.R. Sentence Reading Test ; in addition, other reading tests were given to children when they actually joined the centre. The control group was matched with pupils who went to the centre for chronological age, reading quotient (N.F.E.R. Sentence Reading Test), sex, and, as far as possible, intelligence test score and school ; and this matched group which attended the centre will be referred to as the experimental group. While the matching procedure was not perfect, it can be said with certainty that the teachers of the control group had no idea whatever, that some years later these children would be followed up and retested. Moreover, these fifty children in the experimental group were drawn from fifteen schools so that the classes out of which they were taken were but little reduced in size. At follow-up, then, the control group was also traced and the two reading tests given. See Table 6.

III.—RESULTS.

The data obtained have been given in tables for the sake of clarity.

TABLE 1

DATA RELATING TO THE PERIOD OF REMEDIAL TEACHING.

Variable	Backward Group N=164	Non-Backward Group N=76
I.Q.	$M=98{\cdot}8$, $\sigma=12{\cdot}8$	$M=111{\cdot}3$, $\sigma=13{\cdot}3$
R.A. (in years) at time of referral	$M=6{\cdot}4$	$M=8{\cdot}2$, $\sigma=1{\cdot}0$
C.A. (in years) at time of referral	$M=9{\cdot}2$, $\sigma={\cdot}9$	$M=8{\cdot}7$, $\sigma={\cdot}7$
Mean Reading Quotient $\left(\frac{RA}{CA}\right)$	70	94
R.A. (in years) at end of remedial teaching	$M=8{\cdot}4$, $\sigma=1{\cdot}4$	$M=10{\cdot}0$, $\sigma=1{\cdot}3$
C.A. (in years) at end of remedial teaching	$M=10{\cdot}4$, $\sigma=1{\cdot}0$	$M=9{\cdot}6$, $\sigma={\cdot}8$
Mean Reading Quotient	81	104
Mean length of period of remedial teaching (years)	$M=1{\cdot}2$	$M={\cdot}9$
Extremes (in years) in period of remedial teaching	$0{\cdot}5-3{\cdot}7$	$0{\cdot}25-2{\cdot}9$
Mean improvement in R.A. during remedial teaching (in years)	2·0	1·8
Extremes (in years) of improvement during remedial teaching	Nil—5·4	$\cdot5-4{\cdot}0$

TABLE 2

DATA RELATING TO THE PERIOD SINCE REMEDIAL TEACHING CEASED.

Variable	Backward N=164	Non-Backward N=76
R.A. (in years) at follow-up (Schonell R1)	M=10·0, σ =1·8	11·4, σ =1·2
R.A. (in years) at follow-up (Vernon Graded Word)	M= 9·8, σ =2·0	11·4, σ =1·5
C.A. (in years) at follow-up	M=13·1, σ =1·6	12·1, σ =1·3
Reading Quotient (R.A. from Schonell R1)	76	94
Mean progress (in years) since remedial teaching ceased	1·6	1·4
Mean length of time in years between end of remedial teaching and follow-up	2·7	2·5
Extremes of progress made (in years) since remedial teaching ended :		
12-23 months previously	−0·8 to 4·4 M=1·2 N=64	−0·5 to 2·4 M=0·7 N=28
24-35 months previously	−0·4 to 4·5 M=1·7 N=42	−0·6 to 4·1 M=1·8 N=30
36-47 months previously	0·5 to 3·2 M=2·0 N=31	1·0 to 3·8 M=2·2 N=12
48+ months previously	−1·8 to 5·8 M=1·9 N=27	0·4 to 2·3 M=1·1 N=6

TABLE 3

DISTRIBUTION OF READING AGES (SCHONELL R.1) AT FOLLOW-UP.

	Reading Age in Years										
	−5·4	5·5—6·4	6·5—7·4	7·5—8·4	8·5—9·4	9·5—10·4	10·5—11·4	11·5—12·4	12·5—13·4	13·5—14·4	
Backward	0	4	10	24	28	26	35	24	11	2	164
Non-Backward	0	0	0	1	4	6	28	24	8	5	76

TABLE 4

PERCENTAGE NUMBERS OF CHILDREN GIVING CORRECT REPLIES TO SELECTED QUESTIONS IN THE TEST OF ARITHMETIC CONCEPTS.

Question	Backward	Non-Backward
1.—In the number 26 what does the figure 6 stand for (worth) ? Or, What does it tell us ?	37	68
2.—In the number 26 what does the figure 2 stand for ? Or, What does it tell us ?	44	66
3.—In the number 404, what does the middle figure 0 tell you ?	29	39
4.—In the number $\overset{*}{4}0\overset{\dagger}{4}$ how many times bigger is the 4 with a star than the 4 with a cross ?	10	13
5.—What number is half way between 26 and 32 ? . .	39	61
6.—Write down the answer to this exercise without actually multiplying the numbers : 9×57 divided by 9	13	22
6a.—Explain how you obtained the answer you have just given (in 6). Or, How did you get that ? . .	12	18

TABLE 5

PERCENTAGE NUMBERS OF CHILDREN IN GIVEN STREAMS OF MULTI-STREAM PRIMARY, SECONDARY MODERN OR ALL AGE SCHOOL AT FOLLOW-UP.

Group		Multi-Stream School			
	Single Stream School	Bottom Stream	Next to Bottom Stream in 3 or more Stream School	Top Stream	Inter-mediate Stream
Backward . .	16	59	18	5	2
Non-Backward	20	33	13	14	20

TABLE 6

DATA RELATING TO THE EXPERIMENTAL AND CONTROL GROUPS—50 IN EACH GROUP.

	Experimental Group	Control Group
I.Q.	$M = 116$, $\sigma = 10 \cdot 0$	$M = 112$, $\sigma = 10 \cdot 6$
C.A. (in years) when Experimental group first joined the Centre....	$M = 7 \cdot 9$	$M = 7 \cdot 9$
Reading Quotient (N.F.E.R. Sentence Reading)	$M = 85$	$M = 85$
Length of time (in years) at Centre	$M = 0 \cdot 9$	
R.A. (in years) at follow-up (Schonell R1)	$M = 10 \cdot 7$, $\sigma = 1 \cdot 2$	$M = 10 \cdot 8$, $\sigma = 1 \cdot 3$
R.A. (in years) at follow-up (Vernon)	$M = 10 \cdot 5$, $\sigma = 1 \cdot 5$	$M = 10 \cdot 7$, $\sigma = 1 \cdot 5$
C.A. (in years) at follow-up......	$M = 11 \cdot 3$	$M = 11 \cdot 3$
Reading Quotient (Schonell R1) . .	94	96
Mean length of period (in years) since the experimental group first joined the Centre	$M = 3 \cdot 4$	$M = 3 \cdot 4$

TABLE 7

INTER-RELATIONSHIPS BETWEEN RELEVANT VARIABLES USING KENDALL'S TAU CO-EFFICIENT.

Relationship	Backward Group	Non-Backward Group
I.Q. and progress made during remedial teaching	·22**	·15*
I.Q. and R.A. (Schonell R.1) at follow-up	·27**	·10
Length of remedial teaching and progress made during remedial period	·21**	·35**
Progress made since remedial teaching ceased and time elapsed since	·27**	·30**
Progress made during remedial teaching and R.A. at follow-up	·37**	·24**
I.Q. and progress during remedial teaching holding constant the length of remedial teaching	·20†	·16†

*—Indicates significance at 5 per cent. level.

**—Indicates significance at 1 per cent. level.

†—It is not possible to test the significance of $\tau_{xy.z}$

IV.—DISCUSSION.

The results may be conveniently discussed under a number of headings :

(1) The data relating to the backward group (which has a mean I.Q. close to 100) during the period of remedial teaching shows much the same general trends as the corresponding data in the earlier study (Lovell *et al. op cit.*). There is a rise in the reading quotients of the same order of magnitude in both studies, and overall, we may say that as far as reading ages are concerned, the children in the earlier study, who had had remedial teaching for only one period a week, made equally as much progress per unit time as those now under discussion. Furthermore, the progress made in reading per unit time since remedial teaching ceased is almost identical in both investigations.

(2) There is tendency in the previous investigation for the reading quotients to fall between the end of treatment and follow-up. The same trend is clearly noticeable in both backward and non-backward groups in this study.

(3) Table 3 shows that only some 40 per cent. of the backward group had an understanding of place value involving tens and units at follow-up, and far fewer could fully appreciate place value in relation to hundreds, tens and units. This is a serious matter since an understanding of place value is fundamental in our number system.

(4) Over three-quarters of the children who were backward at time of joining the centres, were found at follow-up, in the bottom or next to bottom stream of a three or more stream school. A further 16 per cent. were in single stream schools, and only 5 per cent. in a top stream. This clearly confirms the view expressed in the earlier paper—that it is only very few children, backward at 8½–9 years of age, who are later found among the academically abler pupils. Not a single child of secondary school age in this group was found in a grammar school at follow-up.

(5) The findings given in Table 6 came somewhat as a surprise. After a mean period of almost $3\frac{1}{2}$ years from the time of referral, no difference in the mean reading ages could be found between those who did, and those who did not attend the remedial centres. While the matching procedure could not be perfect, it was done with great care. Moreover, those who went to the centres were regarded as the most promising pupils. It is still possible, of course, that up to one year after referral, the mean reading age of those attending the centres might have been ahead of the mean reading age of those in school.

(6) The non-backward group (mean I.Q. 111) made about the same progress per unit time in reading as the backward group during the period of remedial teaching. The reading quotient rose but fell markedly again in the period after remedial teaching ended so that at follow-up the quotient was roughly the same as at the time of referral.

(7) In the arithmetic test some two-thirds of the non-backward group understood place value in relation to tens and units, but the more difficult questions relating to hundreds, tens and units remain poorly answered.

(8) In spite of the high mean I.Q. of the non-backward group, nearly half of them were found in the bottom stream, or next to bottom stream of a three or more stream school at follow-up. One in five of the group was found in a single stream school, and roughly one in seven in the top stream of a multi-stream school. But of those of secondary school age at follow-up, not a single child was found in a grammar school.

(9) It must be stressed that although the progress made by the majority of the children involved in this study was limited, the progress made was of great importance for the pupils concerned. To attain a reading age of 9+ years is a very real progress for the individual.

(10) The correlation coefficient between I.Q. and progress made during treatment (Table 7) is small but significant in the backward and non-backward group. This clearly parallels the finding in the earlier study.

(11) The previous study and this one, which together involved the follow-up of five hundred children, compared children taught singly, and in small groups, for one period per week, with children of somewhat higher I.Q. taught full-time in remedial centres. Whatever arrangements we have, the long-term results seem to be much the same. This surely suggests that we know very little as yet about the aetiology of scholastic failure and indicates the great need for research.

ACKNOWLEDGEMENTS.—Our thanks are due to the Chief Education Officer and the Senior Educational Psychologist of the Local Education Authority concerned for their generous help in conducting the enquiry and for permission to publish our findings.

V.—REFERENCES.

LOVELL, K., JOHNSON, E., PLATTS, D. (1962). A summary of a study of the reading ages of children who had been given remedial teaching. *Brit. J. Educ. Psychol.*, 32, 66-71.

(*Manuscript received, 10th May,* 1962.)

DEVELOPMENTAL PATTERNS OF SEVERELY SUBNORMAL CHILDREN

BY MARY WOODWARD* AND DIANA J. STERN†
(*Fountain Hospital, London*)

SUMMARY. Eighty-three severely subnormal children aged under 9 years were observed with the aim of comparing their development in different areas and of examining the relation of various aspects of development to Piaget's sensori-motor stages. It was found that locomotor development was significantly in advance of speech development. Children classified at various sensori-motor stages differed significantly in speech and in locomotor development. Certain developments in language, drawing, performance ability and social responses were associated with the attainment of the last sensori-motor stage, thus suggesting that the achievement of the end of the sensori-motor period is an important step in the development of severely subnormal children.

I.—INTRODUCTION.

A PREVIOUS study (Woodward, 1959) investigated a method of classifying the severely subnormal in terms of the six stages of sensori-motor development which have been distinguished by Piaget (1953 and 1955). The results indicated that severely subnormal children pass through these steps in the same sequence as Piaget suggests for normal children up to the age of two years. The aim of the present enquiry was (i) to determine the locomotor, language and social development of children classified at these sensori-motor stages, and (ii) to examine patterns of development in young severely subnormal children.

The main theoretical interest attaches to the comparison of language development and the Piaget stages. At the last sensori-motor stage, children solve problems in such a way as to suggest that they are able to anticipate the future position of an object before they move it, and they persistently search for hidden objects in a manner which suggests that they are aware of the permanence of objects. According to Piaget (1950), this behaviour indicates that the child has some means of representing objects to himself in their absence, and this implies the development of ' a symbolic function ' which makes possible the systematic acquisition of language. This suggests the hypothesis that severely subnormal children will make a similar development in language at the end of the sensori-motor period.

II.—SUBJECTS.

The subjects were resident at the Fountain Hospital, London. The total hospital population born during 1951 or later was selected in the first instance as the group to be investigated. Children were then excluded from this group if they had any condition, revealed in the medical examination, such as cerebral palsy, gross hydrocephalus, or a visual or auditory defect, that is likely to retard any aspect of development. In addition, two children with I.Qs. of over 50 were omitted from the study. This left a group of fifty-four children. During the following eighteen months twenty-nine new admissions were added to the group ; these were consecutive cases without motor or sensory disabilities. This yielded a total group of eighty-three children (fifty-one boys and thirty-two girls), aged 11 months to 8 years 7 months (mean age=4·85 years). Three children had Terman-Merrill I.Qs. of 42, 41 and 31 ; the rest failed one or more of the six two-year items on this scale (forty-nine failed all six), thus falling below the lower limit of the scale.

---

* Now at University College, Swansea.

† Now at Child Guidance Clinic, Bexleyheath, Kent.

III.—METHOD.

The main classification was made in terms of Piaget's six sensori-motor stages. Since these stages have been described in detail in the earlier study already referred to, only the main criteria are given below.

Stage I (birth to 4/5 weeks).

Use and extension of reflex activities.

Stage II (1 month to 4/5 months).

Integration of reflex activities with higher activities, e.g., co-ordination of vision and hearing and vision and grasping.

Stage III (4/5 months to 8/9 months).

Repetition in the same form of chance-made movements that produced an effect on objects that the child saw or heard (e.g., shaking a rattle). Following the line of fall when an object is dropped and recovering half-hidden objects.

Stage IV (8/9 months to 11/12 months).

At the beginning of the stage, removing an obstacle in order to grasp a toy which is visible : at the end, removing an obstacle when the object is hidden but searching only in one place. Visual examination of objects by turning them round and round and holding them at various distances.

Stage V (11/12 months to approximately 18 months).

Obtaining an object that is out of reach by pulling a string attached to it, or the support on which it stands ; searching briefly in different places when an object is hidden. Exploratory manipulations which vary the effect on objects, e.g., banging a toy successively on different-sounding surfaces.

Stage VI (approximately 18 months onwards).

Inferring the result of an action before the object is moved, e.g., using a stick in order to obtain a toy that is out of reach ; searching for objects for a longer period (i.e., awareness of the permanence of objects is acheived).

More advanced manipulation was investigated by the following performance tasks from the Revised Stanford-Binet (Terman and Merrill, 1937) and Merrill-Palmer (Stutsman, 1931) scales of intelligence :

Four-brick tower.
Three-figure form board.
Three-figure form board rotated.
Wall in peg boards A and B.
Block bridge.
Bead threading.
Seguin form board.
Little pink tower.
Six-cube pyramid.

If a child did not succeed with any of the first four tasks, the rest were not given.

It was subsequently found necessary to derive a performance score for each child. One point was awarded for each all-or-none task that was passed, according to Merrill-Palmer and Stanford-Binet criteria of success ; timed tasks (e.g., peg boards and Seguin form board) were given weighted scores on the basis of speed. Refusals were allowed for as in the Merrill-Palmer method of scoring (Stutsman, 1931).

The subjects were in the phase of development which, in normal children, would be assessed by means of developmental scales. These were used in the present enquiry as a basis for the assessment of speech and locomotor development, but experience of administering such scales to severely subnormal children suggested certain modifications.

The locomotor section of Gesell's schedule (Gesell, 1940 and 1941) was used, with the exception of items for which special equipment, such as walking boards, was needed. Gesell's instructions for administration were followed. In cases of doubt additional information was obtained from nursing staff. At the more advanced levels, 2½ years onwards, some items (e.g., standing on one foot) require the child to imitate the examiner. The co-operation of withdrawn children could not be obtained for these items, and teaching and nursing staff could not supply information on them. Thus, these imitative items were omitted from the scoring for all children.

The language sections of developmental scales include, in addition to vocalisation and comprehension items, others relating to response to sound, manipulation of a bell and social responses. When emotionally disturbed children obtain low scores on language scales because they fail to respond to sounds, toys and people, the result does not accurately describe their language development. In this study only items of vocalisation and comprehension were included. Severely subnormal children frequently show scatter on a language scale, verbal comprehension being relatively more advanced than vocalisations. Thus, these two aspects of language were treated separately.

For speech development the vocalisation items were abstracted from Gesell's scale, together with the age norms assigned to them by him. Scores in terms of 'speech ages' were based on these norms. Records were made of the type of vocalisations uttered by the child during the examination, and further information on habitual sounds was obtained from nursing and teaching staff. Some children made no other sounds except grunts and squeaks, and, although these occur as items in Gesell's scale, speech ages were not assigned on the basis of these alone, so that these children were omitted from the analysis of the results on speech. The amount of the child's spontaneous vocalisation during the examination was rated on a three-point scale.

The following items were used for the investigation of comprehension :

Response to a gestured instruction.
Response to simple commands in a concrete situation (e.g., " Shut the door, sit down, stand up ").
Comprehension of 'give' in relation to a real common object (e.g., cup, spoon, ball).
Comprehension of 'show' in relation to a real common object.
Response to the question : " Would you like a sweet ? "
Response to the statement : " There's a sweet in the box."
The simple commands item No. 6 of year II of the Terman-Merrill scale (Form L).

Data collected on social development included the child's social response to the examiner and to a familiar nurse, his response to difficulty, and any attempts at communication or at drawing the examiner's attention to his activities. Information on the achievement of feeding skills was obtained from nursing staff.

IV.—RESULTS.

(*a*) *Performance ability.*

The number of children falling in each of Piaget's sensori-motor stages is shown in Table 1, together with the average age of the children in each stage.

TABLE 1

SENSORI-MOTOR STAGE AND LOCOMOTOR AGE.

Sensori-motor Stage	N.	Mean C.A. (yrs.)	Locomotor age in years		
			Basal age Mean	Highest point	
				Mean	Range
III	10	2·48	0·43	0·60	0·38—0·77
IV	10	3·80	0·83	1·02	0·54—2·0
V	29	4·83	1·41	1·55	0·77—3·0
VI	34	5·88	2·32	2·38	0·94—4·0

Twenty-five of the thirty-four Stage VI children succeeded with two or more of the performance tasks, their scores ranging from 1 to 21, but no child below Stage VI succeeded with any of them, a difference which is significant at beyond the 0·001 level ($\chi^2=35{\cdot}4$, d.f.$=1$). Of the nine Stage VI children who failed all these items, five showed signs of severe behaviour disorder in the form of visual avoidance, or prolonged distress. The other four were the youngest at this stage, aged 3 years to 4 years 4 months.

(*b*) *Locomotor development.*

Two locomotor ages for each child derived from the Gesell scale as follows:

(1) The age level at which the child succeeded with all the items (basal age).

(2) The highest point at which a success was achieved.

The locomotor scores of children in each sensori-motor stage are shown in Table 1. From this it can be seen that the average locomotor age increases with successive sensori-motor stages, but there is a wide range at each stage. Nevertheless, the difference between the means for the highest point score of Stages V and VI are statistically significant ($t=8{\cdot}42$, $p<0{\cdot}001$, d.f.$=61$). In view of the small numbers in Stages III and IV and of the small differences between the means, these two stages were combined; the resulting average is significantly lower than the Stage V average ($t=4{\cdot}75$, $p<0{\cdot}001$, d.f.$=47$). In no instance was a child's locomotor age (highest point) level with his chronological age.

A difference between the basal age and the highest point occurred in thirty-six cases. Among the children who could not walk without support there was a consistent pattern in the type of item which accounted for the scatter. The details are given in Table 2 for the twenty-one children whose upper limit (highest point) was below one year.

TABLE 2

ANALYSIS OF ITEMS PASSED AND FAILED BY TWENTY-ONE CHILDREN SHOWING SCATTER ON THE GESELL LOCOMOTOR SCALE.

Highest point of locomotor development	N.	Type of item passed (+) and passed (−) at highest point				Range of scatter between last standing item + and last sitting item +*
		Sitting items		Standing items		
		+	−	+	−	
48 w	7	7	0	0	7	8 to 24 weeks
40 w	5	4	1	0	5	12 to 28 weeks
32–30 w	4	4	0	1	3	4 to 8 weeks
20–28 w	5	4	1	—	—	No standing items at these ages.
	21	19	2	1	15	

* Age level at which last success achieved for a sitting item subtracted from age level at which last success is achieved for a standing item.

It will be noted that in nineteen cases the success at the highest point was a sitting item, and in one case only it was a standing item, and that a sitting item was failed at the highest point in only one case. Standing items, on the other hand, were failed at the highest point by fifteen of the sixteen children whose highest point was thirty-two weeks and over (there are no standing items below the age of thirty-two weeks). Thus, the scatter in the majority of cases was due to failure to show various standing responses over a range of the scale at which sitting responses were present. New admissions, who had recently come from home did not differ significantly from those who were already resident in hospital either in locomotor score or in the relative number showing a scatter over the locomotor scale.

Locomotor score (highest point) was positively correlated with performance score among the Stage VI children (rho=0·68, p=0·001). Hence, this factor was controlled in group comparisons, as was chronological age also, though a significant correlation between age and locomotor score (highest point) within any stage was found only for the group at Stage VI with performance scores of 0 (rho=0·625).

In view of the tendency of Mongols to lower muscle tone (Hilliard and Kirman, 1957), their locomotor development was compared with that of non-Mongols. Twenty-three pairs were matched for age, Piaget stage and performance score, the average ages of the two groups being 4·96 years and 5·025 years, respectively. The mean locomotor age (highest point) of the Mongols which was 1·76 years (S.D.=0·98) differed very little from that of the non-Mongols (M=1·77, S.D.=0·96). The basal age means were similarly close (1·63 years and 1·64 years).

(c) *Speech development.*

In fifteen cases no sounds (apart from squeaks, crying and laughing) were heard during the examination, and the nursing staff similarly reported having heard no other sounds from these children. The 'speech age' of the remaining sixty-eight children is shown in Table 3.

TABLE 3

LOCOMOTOR AND SPEECH DEVELOPMENT.

Sensori-motor Stage	N	Locomotor age Highest point Mean (years)	Speech Age	
			Mean (years)	Range (years)
III	6	0·617	0·487	0·23—0·77
IV	8	1·058	0·403	0·15—0·69
V	21	1·660	0·784	0·54—1·25
VI	33	2·330	1·196	0·62—2·0

The first interesting feature of this table is the wide range in speech age at any Piaget stage ; the second is the increase in the upper limit of speech development with successive stages. At Stages III and IV the most complex sound made by any child was a four-syllable babble ; jargon, interwoven with single words, was the most advanced development at Stage V. Two Stage V children produced two-word combinations in an echolalic manner, but meaningful and relevant word combinations and short sentences occurred only in the speech of some of the Stage VI children.

The difference between the means of the Stage V and Stage VI children is statistically significant at beyond the 0·001 level ($t=3{\cdot}92$ for 52 d.f.). The Stage V average is also significantly different from the combined mean of Stages IV and III ($t=3{\cdot}52$, $0{\cdot}01>p>0{\cdot}001$, d.f.$=33$).

In view of the mutism that is sometimes associated with severe behaviour disorders, it was postulated that the absence of speech sounds, or very retarded speech development, would be related to behaviour disturbance in children at Stages V and VI, among whom absence of speech could scarcely be attributed merely to extremely slow development. Very retarded speech was defined as up to and including twenty-eight weeks (vowels only) for Stage V and up to and including thirty-six weeks (two-syllable babble) for Stage VI. Behaviour disorder was defined by marked visual avoidance or prolonged distress when approached, or mannerisms which warded off social contact. Eight of the eighteen children at Stages V and VI who were extremely retarded in speech development, or who made no sounds, showed signs of behaviour disorder, compared with three of the remaining forty-five children at these stages ; this difference is significant between the 0·01 and 0·001 levels ($\chi^2=10{\cdot}25$, with Yate's correction applied, d.f.$=1$).

Among the children at Stage VI, speech development was positively correlated with performance score when the four children with severe behaviour disorders were omitted (rho$=0{\cdot}38$, $p=0{\cdot}05$). Rank order correlations of age and speech development within each Piaget stage and performance score group were not significantly different from zero in any instance.

The bearing of the period of residence in hospital on speech development was examined by comparing the speech development of twenty-three new admissions and twenty-three others matched for age, Piaget stage and performance score where applicable. The average speech ages of the two groups were 0·83 years and 0·81 years, respectively. (When new admissions were excluded, there were too few children left at any stage for length of stay in hospital to be correlated with speech age.)

Twenty-two pairs of Mongols and non-Mongols were matched for age, Piaget stage and performance score. A mean difference of 0·126 years in speech age, favouring the Mongols, was found, but there were six children with severe behaviour disorders in the non-Mongol group and non among the Mongols. Excluding these six children and their Mongol pairs, the mean speech ages for the remaining sixteen pairs were 1·028 years (S.D.=0·405) for the Mongols and 0·896 (S.D.=0·5) for the non-Mongols; this difference was not statistically significant.

A comparison of speech and locomotor development in the same children within each Piaget stage is made in Table 3, excluding the children who made no speech sounds. It will be seen that, for all groups, the average locomotor age is considerably in advance of the average speech age.

For the whole group of sixty-eight children, the average speech age (0·92 years) was significantly lower than the average locomotor age (1·82 years) for the highest point score. (t for correlated means=9·82, $p < 0 \cdot 001$, d.f.=67).

In fifty-five of the sixty-eight cases (81 per cent.), both the highest point of locomotor development and the basal age were higher than the speech age. In four cases one or both was the same as the speech age; both locomotor ages were lower than the speech age in four cases, and in the remaining five cases only the basal age was lower.

(*d*) *Verbal comprehension.*

None of the children classified at Stages III and IV showed any understanding of speech, nor did twenty-five of those at Stage V, nine of whom responded to instructions given by gesture. The remaining four Stage V children gave evidence of verbal comprehension in simple concrete situations. All the children classified at Stage VI, on the other hand, showed evidence of some verbal comprehension. Fourteen of them responded only at the level of simple concrete situations; of these nine failed the other items listed in Section III, but the other five were too disturbed for these to be administered. The remaining twenty children succeeded with three or more of the items listed in Section III (including concrete situations and excluding gestures). Comparing Stages V and VI for the number of children in three categories of verbal comprehension (1) success with three or more items, (2) verbal comprehension in concrete situations only, and (3) none at all, the differences are significant at beyond the 0·001 level ($\chi^2=50 \cdot 36$, d.f.=2).

The relation between verbal comprehension and Piaget's sensori-motor stages was further investigated by an examination of clinical records, which contained the Piaget stage and the results of attempts to administer the Terman-Merrill revision of the Binet scale. Children who succeeded with one or more of the year II verbal items were invariably at sensori-motor Stage VI.

(*e*) *Drawing.*

The drawing ability of children at the various Piaget stages is shown in Table 4.

TABLE 4

SENSORI-MOTOR STAGE AND DRAWING.

Drawing	Sensori-motor stage			
	III	IV	V	VI
Nothing	10	10	20	0
Faint marks or dots	—	—	7	3
Straight scribble	—	—	2	13
Circular scribble	—	—	—	3
Copy of lines	—	—	—	6
Copy of circle, partial*	—	—	—	6
Copy of circle, correct	—	—	—	2
Copy of Cross	—	—	—	0
Totals	10	10	29	33

* By circular scribble.

It will be seen that the use of a pencil begins to occur at Stage V, but extends to vigorous scribbling in only two cases. All but three of the Stage VI children, on the other hand, scribbled freely or copied models of straight lines and circles. (One refused to handle the pencil.) The differences between Stages V and VI, in three drawing categories ((1) none, (2) marks or dots, and (3) scribbles or copies) are significant at beyond the 0·001 level ($x^2=28\cdot1$, d.f.$=2$).

(*f*) *Feeding skills.*

Feeding skills achieved by children at various Piaget stages are shown in Table 5.

TABLE 5

SENSORI-MOTOR STAGE AND FEEDING SKILLS.

	Sensori-motor Stage			
	III	IV	V	VI
Cup: Does not drink	10	7	7	—
Drinks if cup held	—	3	8	—
Helps to hold cup	—	—	5	2
Holds cup without help	—	—	9	32
Does not feed self with fingers	10	7	13	—
Feeds self with fingers	—	3	16	34
Spoon: Does not use	10	10	23	2
Uses and spills	—	—	5	6
Uses without spilling	—	—	1	26
Total in each stage	10	10	29	34

(g) *Social responses.*

All the children in the group showed awareness of the investigator and a nurse or teacher, most by looking at the investigator when she approached; the social awareness of some was indicated in a negative way by visual avoidance. Others were distressed in the presence of the investigator, or began a mannerism or slapped their faces when approached.

The response of smiling was found at all stages, though not all the children in a given stage showed it; the incidence was naturally affected by the behaviour described in the previous paragraph. Few children made a vocal response when talked to by the investigator, but spontaneous vocalisations occurred in more cases. A rating of 1 (few or no spontaneous vocalisations) compared with a rating of 2 or 3, was given to significantly more of the children who showed signs of behaviour disorder than to the rest ($\chi^2 = 20{\cdot}02$, p. $< 0{\cdot}001$, d.f. $= 2$.).

Differences in the amount of vocalisation (assessed by the ratings) between new admissions and children already resident in hospital were not statistically significant.

Communication, in terms of imparting information or making a demand, was observed mainly among the Stage VI children, twenty-two of the thirty-four pointing for this purpose, with or without a vocal accompaniment. Only two Stage V children communicated, and none below this level did so.

Other social responses were observed only among the Stage VI children. Seventeen (50 per cent.) of them drew the investigator's attention to their activities, or kept looking to make sure the investigator was watching their performance; and twelve children sought the investigator's help when in difficulty.

V.—DISCUSSION AND CONCLUSIONS.

There are two main questions arising from these findings that call for discussion. One concerns the pattern of development in severely subnormal children, and the other is the relation of other aspects of development to Piaget's sensori-motor stages.

The group is too deviant for locomotor and speech scores to be equated in terms of standard scores (in any case Gesell does not provide the necessary data for this). The expression of the results in terms of 'locomotor ages' and 'speech ages' is probably inexact, and consequently comparisons of rates of development in different aspects are approximate. Nevertheless, the difference between the locomotor score and the speech score of the whole group is so large that it would appear that the severely subnormal, although grossly retarded in all areas of development, tend to develop much more rapidly in the locomotor sphere than they do in the acquisition of speech. If the approximate ages suggested by Piaget for the sensori-motor stages are accepted, then at all stages (with the possible exception of III) locomotor development appears to be in advance of the development of sensori-motor intelligence, which, in turn, is in advance of the development of speech.

The twin study of Gesell and Thompson (1929) indicates that locomotor development is mainly influenced by maturation, since intensive training made little difference to subsequent locomotor scores. Much of Piaget's account of the development of sensori-motor intelligence is a description of a learning process, and learning obviously enters into the acquisition of speech. Thus, it can be postulated from these results that the severely subnormal develop relatively more rapidly in the aspect which is mainly maturational, and that they develop relatively more slowly in aspects which depend more on learning.

Another finding on locomotor development which calls for further investigation is that of a large scatter over the locomotor scale by the majority of the children who could not walk. This suggests that steps in the sequence of behaviour that lead to the attainment of the upright posture do not, in the severely subnormal, correspond with steps in the development of the sitting posture as in normal children. One possible psychological factor accounting for this is a difference in the amount of individual attention and encouragement to stand received by children in hospital and at home. New admissions did not, however, differ significantly from the children who had been resident in the hospital for a longer period, either in locomotor score, or in scatter over the locomotor scale. Although children with cerebral palsy were excluded from the study, it is, of course, possible that there were motor lesions which were not observable clinically. Comparison of these data with post-mortem findings would be valuable.

The limited verbal stimulation of children in hospital might also be put forward to account for the finding of greater retardation in speech development, relative to locomotor development and the sensori-motor stages. The results of a number of studies (reviewed by McCarthy, 1954) have indicated that normal children living in institutions develop language more slowly than children living at home. A similar finding in respect of severely subnormal children has been reported by Lyle (1960). However, in the present study new admissions did not differ in average speech score from children already resident in hospital. This finding is not necessarily at variance with Lyle's, since he investigated older children whose language was more advanced. They could identify and name objects, whereas the present study was mainly of pre-verbal utterances. It is thus possible that different types of environment become more important once words are beginning to be systematically acquired. The studies of normal children, however, give conflicting results as to the age at which an unstimulating environment begins to affect language development, some reporting an effect on phonemes in the first few months and others finding a difference only after the age of 2 years (McCarthy, 1954). The hypothesis that an unstimulating environment influences language development only after the beginning of the systematic use of speech might also account for the different results in respect of Mongols in this study and in Lyle's. Lyle (1960) found that Mongols, whether living in hospital or at home, were more backward in language than comparable non-Mongols. The Mongols of the present group did not differ significantly from non-Mongols.

Further study might usefully be made of the findings of the present study that children with severe behaviour disorders tended to be very retarded in speech development or to make no sounds. This result also indicates that behaviour disorder is a variable that should be controlled in comparative studies of language development.

The second main question to be discussed is whether children classified into Piaget's sensori-motor stages are clearly differentiated in other aspects of behaviour. For most aspects of development there is little difference between Stages III and IV. This is probably to be explained by the fact that all the children at Stage IV were only at the beginning of this stage. This appears to be a transitional stage, and very few of the severely subnormal children who have been observed have been at the end of the stage. Children are classified at the beginning of Stage IV if they brush aside a transparent screen in order to obtain a toy that it covers, though in the extent of their awareness of the permanence of objects they are the same as those at Stage III (recovering a half-hidden object, but not a completely hidden one).

However, significant differences in both locomotor and speech development were found between successive stages (III and IV together compared with V, and V compared with VI). Children at Stage VI were significantly differentiated from those at Stage V in the development of verbal comprehension. It is also of particular interest that it is only Stage VI children who can succeed with, or even comprehend, the first verbal items of the Revised Stanford-Binet scale. For these it is necessary to identify objects by their names and to carry out simple instructions about objects. This finding supports Piaget's contention that the achievement of Stage VI, which involves a concept of the permanence of objects, is necessary for the acquisition of meaningful speech. Some children in the present group who were at Stage VI did not use meaningful speech, or identify objects by name ; whether these only lacked sufficient stimulation, or whether another factor is involved, as a matter for further investigation.

It was also only Stage VI children who succeeded with any performance items drawn from pre-school scales. Thus, the attainment of Piaget's sensori-motor Stage VI, at any rate in the severely subnormal, appears to be necessary for success with the earliest items, both verbal and non-verbal, of pre-school scales which are applicable to normal children from the age of 1½ to 2 years.

Results for other aspects of behaviour also underline the importance of the attainment of the end of the sensori-motor period in the development of severely subnormal children ; only Stage VI children were able to copy lines and a circle, and with two exceptions vigorous scribbling occurred only at this stage. Communication began to appear at Stage V but was common only at Stage VI, and it was only Stage VI children who sought help in difficulty, and who indicated that they wished their activities and achievements to be noticed by an adult.

To turn to the practical implications of the findings, the first application is in clinical practice. All the Stage VI children in this study showed evidence of some verbal comprehension. Hence, if a child at Stage VI is found to have no verbal comprehension, then further investigation is indicated, particularly of the possibility of deafness.

The results of this study also suggest when it is useful for nursing staff to spend time encouraging a child to use a cup and spoon. The findings indicate that children who have reached Stage VI can certainly be taught to feed themselves, and they also suggest that it is worthwhile to persist with Stage V children. However, the attempt is probably fruitless until the child reaches these two stages.

In the educational field the results indicate that children who have attained the last sensori-motor stage would benefit from attending a training centre, since new developments in language, drawing and performance ability occur during this stage, which appears from this study to be a major event in the development of severely subnormal children.

VI.—REFERENCES.

GESELL, A. (1940). *The First Five Years of Life.* New York : Harper.
GESELL, A., and THOMPSON, H. (1929). Learning and growth in identical infant twins. *Genet. Psychol. Monag.*, 6, 1-124.
GESELL, A., and AMATRUDA, C. S. (1941). *Developmental Diagnosis.* New York and London : Hoeber.

HILLIARD, L. T., and KIRMAN, B. H. (1957). *Mental Deficiency*. London : Churchill.

LYLE, J. (1960). The effect of an institution environment upon the mental development of imbecile children : II—Speech and language. *J. Ment. Defic. Res.*, 4, 1-13.

MCCARTHY, D. (1954). Language development in children. Carmichael, L. (Ed.). *Manual of Child Psychology*. New York : John Wiley.

PIAGET, J. (1950). *The Psychology of Intelligence*. Trans. by Piercy, M., and Berlyle, D.E. London : Routledge and Kegan Paul.

PIAGET, J. (1953). *The Origin of Intelligence in the Child*. Trans. by Cook, M. London : Routledge and Kegan Paul.

PIAGET, J. (1955). *The Child's Construction of Reality*. Trans. by Cook, M. London : Routledge and Kegan Paul.

STUTSMAN, R. (1931). *Mental Measurement of Pre-School Children*. Yonkers, New York : World Book Co.

TERMAN, L. M., and MERRILL, M. A. (1937). *Measuring Intelligence*. London : Harrap.

WOODWARD, M. (1959). The behaviour of idiots interpreted by Piaget's theory of sensorimotor development. *Brit. J. Educ. Psychol.*, 29, 60-71.

(*Manuscript received*, 17*th May*, 1962.)

27.8.70
9225

ARE RURAL CHILDREN HANDICAPPED BY THE USE OF SPEEDED TESTS IN SELECTION PROCEDURES?

BY

C. ANNE MORETON

(*Charlotte Mason Training College*)

AND

H. J. BUTCHER

(*Department of Education, University of Manchester*)

SUMMARY. It is widely believed in country areas that the country child is at a disadvantage in the working of speeded tests. This hypothesis was tested by the comparison of two samples, each of about 300 children aged 10–11, one from rural Westmorland, one from central Manchester. Power and speed scores were obtained from N.F.E.R. tests of Arithmetic, English and Verbal Reasoning. Almost all net differences were statistically significant in the expected direction, tending to confirm the hypothesis, but were not sufficiently large to suggest a need for revision of selection procedures.

I.—INTRODUCTION.

IN the course both of teaching in rural schools and also of supervising school practice in a rural area, one of the authors has frequently met with complaints from teachers and parents that the country child is at a disadvantage in the working of speeded tests used in selection procedures. This disadvantage is often attributed to the slower tempo of life in the country as compared with that in urban areas.

The present research is confined to testing the hypothesis that such a disadvantage exists, without any detailed exploration of possible reasons or correlates. But it is believed to be the first systematic enquiry of its kind and may be timely in view of recent renewed interest in the concepts of speed and power in mental tasks (e.g., Furneaux, 1960).

II.—PREVIOUS RESEARCH.

The literature on mental 'speed' and 'power' is extensive, and often conflicting. About 150 references are given by Morrison (1960) and it is, therefore, outside the scope of this paper to offer anything like a comprehensive survey. One of the difficulties in this field of research is that in practice measures of both speed and power are usually adulterated or relative, since a pure test of speed is of trivial application, and a pure test of power impracticable. Furneaux (1960) is probably correct in holding that such concepts as 'speed,' 'power,' 'accuracy' and 'difficulty' require a new and more operational definition if conflicting results are to be avoided. However, certain trends—one might almost say fashions—in research do appear to emerge.

In the earliest researches there was disagreement between those who maintained that speed of reaction was positively correlated with higher mental processes (Gilbert, 1894, Baguley, 1901) and those who found no such relationship (Binet, 1900; Wissler, 1901; Whipple, 1904). As a result of further research (Burt, 1909; Brown, 1910; Kruger and Spearman, 1914; May, 1921; Bernstein, 1924; Spearman, 1927), the view that such a relationship does exist became generally accepted, although the relative importance and predictive power of these two components of mental ability is still not entirely settled.

Since about 1930, attempts have been made, foreshadowed by Bernstein (1924), to interpret speed of mental work as a personality factor (e.g., Kennedy, 1930; Himmelweit, 1946). In contrast, Eysenck (1953), discussing the work of Furneaux, has maintained that speed is the fundamental component of ability, and that measures of power, which depend on persistence, involve personality factors.

Broadly speaking, however, the orthodox view appears still to be, firstly that speed and power are highly correlated, secondly that power is an important component, if not the more important component of general ability, thirdly, that tests, as used for example in the 11+ selection procedure, achieve an adequate compromise. Thus, Peel (1956), commenting on the kind of criticims that gave rise to the present investigation, concludes both that in general at 11+ " the abler child is the speedier child," and that in any case the time-limit imposed on most group tests in practice is not so stringent as to make speed an over-riding factor.

Among psychologists as well as among the general public there are some who might still query this view (Porebski, 1954; Myers, 1952, 1960; Mollenkopf, 1960). Myers, for example, writes: " In timed selection and achievement tests, we may be defining the purposes of education as including, perhaps emphasising, the teaching of the value of snap judgment and educated guessing, over the value of deliberate careful thought."

In view of the continuing interest in these questions, it is mildly surprising to find how little research has been done into the social correlates of power and speed. Klineberg (1928), Lambeth and Lanier (1933) and Knapp (1960) investigated racial differences, but urban-rural differences appear to be relatively unexplored. The most relevant study known to the writers is that of Smith (1948), whose conclusions about Hebridean children include the following:

> " It is clear in these circumstances that the time factor has a profound effect on test results over and above the normal effect. A rigidly timed test does not seem to discriminate very well in the range of the intelligence scale, because mental processes which are more slowly initiated and pursued than is usual are constantly being interrupted and suspended by the imposed speed of the examiner . . . There is little evidence of the modern tendency to race with time. It is rightly considered infinite and unlikely to interfere with the expected sequence of things; so in the normal course of events there is no need for undue haste. Probably if the children had grasped more fully the significance of the stop-watch in the experiment, the scores would have been more flattering to them."

Smith's research was primarily concerned with bilingual children, and there can be few rural areas comparable in isolation with the Outer Hebrides, yet it seems reasonable to postulate a gradation of natural tempo between such children and urban children. The aim of the present research is to determine whether English rural children in a relatively isolated area are handicapped to some degree in a similar way.

III.—CHOICE OF SAMPLES.

(*a*) *The Manchester sample.*

The Manchester sample of 297 was drawn from six schools in a lower working-class area near the centre of the city. As will be seen from Table 1, the largest number from any one school was 74 and the smallest number 36. In the total sample there were 153 boys and 144 girls. The average age was 10·11.

TABLE 1

School	No. on Roll	Girls	Boys	Total
1	254	17	19	36
2	247	28	32	60
3	228	21	19	40
4	302	40	34	74
5	290	27	34	61
6	271	11	15	26
			N=297	

School No. 1 was a junior mixed school on the edge of a rapidly deteriorating area and, according to the headmistress, has over the past five years lost a number of good working-class families who have moved out to one of the newer estates. Two coloured boys were concerned in the tests, a half-caste and a Nigerian ; many coloured families are now invading this part of Manchester. There were several ' problem ' families, but on the other hand, still many families who had lived in the district for one or two generations.

School No. 2 was a junior and infant school. Again, there was some evidence of deterioration, both to property and living standards in this area, although on the whole, the school had maintained its standards. There were several coloured children in the school who seemed to be at least of average ability and who also seemed to fit in very well with the community.

School No. 3 was a junior school. It was in a similar type of area to that of School No. 2, although here there was more evidence of deterioration of property and living standards. The coloured children did not appear to be of the same calibre as those in the previous school.

School No. 4 was a junior school. This school was in rather a better residential working-class area but the population of the school was a little mixed and there was some slight evidence of a deterioration in the background and living standards of the children over the past few years, although actually the proportion of those selected for grammar schools was above average.

School No. 5 was what is generally known as a slum school. There is a large floating population and many families live in rooms and sub-standard houses. There are about thirty coloured children in the school. It is interesting to note that at this school a coloured boy who took no extra time for any of the tests had both the highest speed raw score and quotient and also the highest time-limit raw score and quotient of the Manchester sample in Arithmetic.

School No. 6 was the nearest to the centre of the city. Some of the children were from very good working-class homes with interested parents, others were from poor homes both economically and socially. There was a small proportion of coloured children in the school and also a variety of nationalities (e.g., Spaniards, Nigerians and Indians) with attendant language difficulties. Considering the prevailing living conditions, the number going forward to either grammar or technical schools or grammar-stream in secondary modern schools was relatively high.

(*b*) *The Westmorland Sample.*

The Westmorland sample of 274 was drawn from all the rural schools in the county except two which had to be withdrawn owing to administrative difficulties. The largest number from any one school was twelve and the smallest number one. There were 142 boys and 132 girls in this sample, with an average age of 10·8. The total number of schools involved was fifty-eight, the largest number on roll being ninety-six, and the smallest six. There were twenty-five one-teacher schools, twenty-seven two-teacher schools, five three-teacher schools and one four-teacher school. It is probably true to say that the majority of the children in these schools were the children of agricultural workers, although there is a slight but steady trend towards an increasing proportion of children whose fathers work in the towns and travel out each day. The majority of such parents are office workers and/or members of the professions.

Considerable care was taken to confine the sample to genuinely rural areas, by the omission of even small towns and of areas that are fast becoming urbanised. The number of old Westmorland family names gives support to the assumption of a strong ' native ' element. There are for example, thirteen of these known to the writers, all of which occur on an average four times. It is not claimed, however, that the two samples are equated for socio-economic status, and it is probable that the Westmorland sample contained a higher proportion of children from middle-class homes.

IV.—METHOD.

(*a*) *The tests used and their administration.*

The tests used were the NFER Arithmetic Test 8, English Test 8, and Verbal Reasoning Test 8a. Considerable practical difficulties were envisaged in conducting a programme of testing over a wide rural area with poor communications and a high proportion of one-teacher schools, in which such a programme can easily disrupt the whole school.

The solution adopted was that the Westmorland results should be based on the Authority's preliminary testing, with some additional instructions to supervisors, and that the same tests should be administered to the Manchester children under as closely similar conditions as possible (but without, of course, forming part of the Manchester Authority's official programme of testing).

Such a solution is not ideal, and might even introduce differences in absolute level of performance, but it was considered unlikely seriously to disturb the *relative* differences between speed and power which form the main topic of investigation.

Instructions to Supervisors.

In addition to the usual instructions for administration issued with the tests as to strict time limit, supervision and marking, the following instructions were sent to each school :

Additional Instructions to Supervisors
Arithmetic Test 8 (50 *Minutes*), *English Test* 8 (50 *Minutes*)
Verbal Reasoning Test 8a (45 *Minutes*)

(1) After each test has been in progress for 10 minutes, instruct the testees to place an 'X' in the margin.

(2) Record time taken by each testee to complete test within the official time-limit. This to be recorded on list containing names and ages (in years and months).

(3) At the earliest convenient time (possibly the following day) allow testees to have back their tests if they are not complete, and to complete them in red pencil. (N.B.—All items omitted on the previous occasion must be completed.)

(*b*) *Types of Score desired.*

The ten-minute score is referred to as a *Time-Limit Score* following Davidson and Carroll's (1945) definition as "the number of items correctly answered within a prescribed time-limit." The score achieved on the normal working of the test, i.e., on the first day and within the 'official' time is referred to as the *Speed Score*. The distinction between 'time-limit score' and 'speed score' is thus one of degree rather than of kind. Finally, the total score including any red pencil entries, is called the *Power Score*. Thus, the experiment was designed to give nine scores in all for each child, consisting of three types of score on each of three tests.

(*c*) *Method of standardisation.*

In order to test our hypothesis, it was necessary to make power scores, time-limit scores and speed scores all comparable. Conversion tables incorporating an age allowance and giving quotients were available in the case of speed scores, but not, of course, in the case of the other two types of score, which were peculiar to the experiment.

The power and time-limit scores were, therefore, scaled to the same mean and standard deviation as the speed scores for the combined sample. It was then possible to convert these into quotients comparable with the speed quotients. A section of one of the conversion tables used is reproduced in Table 2.

TABLE 2

SECTION OF CONVERSION TABLE (ARITHMETIC TEST).
(Ages 10 years—10·11 years)

Stand. Score	Raw Score	Age in years and completed months at time of test												Raw Score
		10·0	10·1	10·2	10·3	10·4	10·5	10·6	10·7	10·8	10·9	10·10	10·11	
−1·723	1	79	78	77	77	76	76	75	75	74	74	73	73	1
−1·678	2	82	81	81	80	80	79	79	78	77	77	76	76	2
−1·571	3	84	84	83	83	82	81	81	80	80	79	79	78	3
−1·588	4	86	86	85	84	84	83	83	82	82	81	81	80	4
—1·544	5	88	87	86	86	85	85	84	84	83	83	82	81	5

V.—RESULTS.

(*a*) *A direct comparison of the two samples on speed, power and time-limit scores.*

Table 3 summarises the comparative performance of the two groups on the three tests and under the three conditions of work. The power and time-limit scores have been made comparable with the speed scores by the procedure already described, which also incorporates an age allowance :

TABLE 3

MEANS AND STANDARD DEVIATIONS—CONVERTED SCORES.

ARITHMETIC							
	MANCHESTER		WESTMORLAND		Mean Diff.	Crit. Ratio	Level of Signif.
	Means	S.D.	Means	S.D.			
Speed	105·32	13·49	104·83	12·21	+0·49	·3571	N/S
Power	104·42	12·9	104·55	12·99	−0·13	−·1199	N/S
Time-limit	106·29	13·71	103·17	13·33	+3·12	2·756	1%
ENGLISH							
	MANCHESTER		WESTMORLAND				
Speed	103·94	13·47	101·82	11·14	+2·12	2·055	5%
Power	103·36	13·11	102·18	12·04	+1·18	1·222	N/S
Time-limit	104·16	13·19	101·87	12·02	+2·29	2·171	5%
VERBAL REASONING							
	MANCHESTER		WESTMORLAND				
Speed	104·44	12·40	102·16	11·78	+2·28	2·255	5%
Power	104·07	12·44	102·58	12·12	+1·49	1·450	N/S
Time-limit	104·34	12·25	100·65	13·17	+3·69	3·426	1%

It can be seen from Table 3 that any significant differences are in favour of Manchester group. Furthermore, all these differences are in respect of the speeded conditions of work (i.e., ' speed ' and ' time-limit '). Thirdly, in each test, the more highly speeded the conditions the greater or more significant the difference in favour of the urban children. This is particularly noticeable in the case of arithmetic, but applies in some degree to the other two tests as well.

(*b*) *A comparison of the two samples on difference scores* (*e.g., power minus speed*).

The results reported in the last section already suggested that the urban children were superior to the rural children to a markedly greater degree in the speeded than in the unspeeded conditions, although absolute differences were small. But a crucial test of the main hypothesis depends on a study of second-order differences, i.e., a comparison of the urban and rural children on power-score minus speed-score.

Accordingly, the difference scores were tabulated and analysed, with results as shown in Table 4 below.

TABLE 4

COMPARISON OF WESTMORLAND AND MANCHESTER CHILDREN ON POWER SCORE MINUS SPEED SCORE.

	Manchester	Westmorland	Difference	Critical Ratio	Significance Level
Arithmetic ..	−0·90	−0·28	−0·62	2·17	·05
English ..	−0·58	0·36	−0·94	3·80	·01
Verbal Reasoning	−0·37	0·42	−0·79	2·89	·01

It can be seen from Table 4 above that all three crucial differences, although small, are in the expected direction and are statistically significant. The size of the difference is largest for English and smallest for arithmetic, although the range is not very great, and should in any case not be interpreted too strictly, since the standardising procedure in the experiment ensured comparability of speed and power scores for any one test, but not necessarily between tests.

A similar comparison in respect of power scores minus time-limit scores is shown below in Table 5 :

TABLE 5

COMPARISON OF WESTMORLAND AND MANCHESTER CHILDREN ON POWER SCORE MINUS TIME-LIMIT SCORE.

	Manchester	Westmorland	Difference	Critical Ratio	Significant Level
Arithmetic ..	−1·87	+1·38	−3·25	5·30	·01
English ..	−0·80	+0·31	−1·11	1·82	n.s.
Verbal Reasoning	−0·27	+1·93	−2·20	4·03	·01

The second-order differences shown in Table 5 are again all in the hypothesised direction, two being highly significant, and the third close to statistical significance. Since the ‘ time-limit ’ scores are formally similar to ‘ speed ’ scores, but based on a shorter period of testing time, it might have been supposed that less conclusive results would be obtained than in the comparison between power scores and speed scores. The indication is, however, that even within the short space of ten minutes the measurement of speeded work operates to some extent in favour of urban children.

VI.—CONCLUSIONS.

1.—Children from a working-class area of central Manchester were compared with children from a rural area of Westmorland on moderately speeded tests of arithmetic, English and verbal reasoning as used in an 11+ selection procedure. Manchester scores were slightly higher on the average in all three papers, two of the three differences being statistically significant.

2.—Scores were obtained after 10 minutes' work on the forty-five minute or fifty-minute tests. Rather larger differences were found in favour of the Manchester children.

3.—When the children were allowed unlimited time to complete the papers, the mean differences between the groups were reduced, and none was significant.

4.—Power-minus-speed scores were calculated for each child. Westmorland children were slightly but significantly higher on all three papers.

5.—The hypothesis that speeded tests to some degree handicap rural children was considered to be confirmed. Relative differences, however, being of the order of one unit of I.Q., were so small as not to suggest any revision of the selection procedure.

VII.—REFERENCES.

BAGULEY, W. C. (1901). On the correlation of mental and motor ability in school children. *Am. J. Psychol.*, 12, 193-205.

BERNSTEIN, E. (1924). Quickness and intelligence. *Brit. J. Psychol. Monogr. Supp.* 3, No. 7.

BINET, A., and HENRY, V. (1896). La psychologie individuelle. *L'Année Psychol.*, 2, 119.

BROWN, W. (1910). Some experimental results in the correlation of mental activities. *Brit. J. Psychol.*, 3, 296-322.

BURT, C. (1909). Experimental tests of general intelligence. *Brit. J. Psychol.*, 3, 94-177.

CARROLL, J. B., and DAVIDSON, W. M. (1945). Speed and level components in time-limit scores : a factor analysis. *Educ. Psychol., Measmt.*, 5, 411-427.

EYSENCK, H. J. (1947). Dimensions of personality, Chapter 4, Section 5. London : Kegan Paul.

EYSENCK, H. J. (1953). *Uses and Abuses of Psychology*. 30-37, Pelican.

FURNEAUX, W. D. (1960). Intellectual abilities and problem-solving behaviour, Chapter 5 in *Handbook of Abnormal Psychology*, H. J. Eysenck, Ed.

GILBERT, T. A. (1894). Researches on the mental and physical development of school children. *Yale Psychology Laboratory*, 240-300.

HIMMELWEIT, H. (1946). Speed and accuracy of work as related to temperament. *Brit. J. Psychol.*, 36, 132-144.

KENNEDY, M. (1930). Speed as a personality trait. *J. Soc. Psychol.*, 1, 286-299.

KNAPP, R. R. (1960). The effects of time limits on the intelligence test performance of Mexican and American subjects. *J. Educ. Psychol.*, 51, 14-20.

KLINEBERG, O. (1928). An experimental study of speed and other factors in "racial" differences. *Arch. Psychol.*, No. 241.

LAMBETH, N., and LANIER, L. H. (1933). Race differences in speed of reaction. *Pedagog. Sem.*, 42, 255-297.

MAY, M. A. (1921). Psychological examination in the U.S. Army. *National Academy of Sciences Memoirs*, 15, 415-420.

MOLLENKOPF, W. G. (1960). Time limits and the behaviour of test takers. *Educ. Psychol. Measmt.*, 20, 223-230.

MORRISON, E. J. (1960). On test variance and the dimensions of the measurement situation. *Educ. Psychol., Measmt.*, 20, 231-250.

MYERS, C. T., (1952). The factorial composition and validity of differently speeded tests. *Psychometrika*, 17, 347-352.

MYERS, C. T. (1960). Introduction to symposium on the effects of time limits on test scores. *Educ. Psychol. Measmt.*, 20, 221-222.

PEEL, E. A. (1956). *The Psychological Basis of Education*, 132-4. Edinburgh : Oliver and Boyd.

POREBSKI, O. (1954). A pyschological and statistical study of speed and power as variables of human ability. *Occup. Psychol.*, 28, 218-229.

SMITH, C. A. (1948). *Mental Testing of Hebridean Children in Gaelic and English.* London : University of London Press.

SPEARMAN, C. (1927). *The Abilities of Man.* London : Macmillan.

SPEARMAN, C., and JONES, W. (1950). *Human Ability.* London : Macmillan.

WISSLER, C. (1901). The correlation of mental and phsycial tests. *Psychol. Monogr.*, 1662.

WHIPPLE, C. M. (1904). Reaction time as a test of mental ability. *Am. J. Psychol.*, 15, 489-498.

(*Manuscript received*, 23*rd May*, 1962.)

SOME CHARACTERISTICS OF CHILDREN WITH TWO TYPES OF SPEECH DISORDER

BY

T. G. CROOKES

(*St. John's Hospital, Aylesbury*)

AND

MARGARET C. L. GREENE

(*Stoke Mandeville Hospital, Aylesbury*)

SUMMARY. An investigation was made of twenty successive cases of developmental speech defects. They were divided according to the type of speech disorder (Greene 1960) into a ' Motor ' group (eight children) and a ' Language ' group (twelve children). In the Motor group, but not the Language, there was evidence of delay in general motor maturation. The two groups were compared on measures of mechanical reading, consistency of laterality and W.I.S.C. subtests. Both groups were very poor at reading, but whereas all the Motor group had inconsistent laterality, only a third (i.e., a normal proportion) of the Language were inconsistent ; on the other hand, the Motor group had a much better knowledge of right and left. The Language group, but not the Motor, had poor scores on the Arithmetic and Digit Span subtests. Also the Language group was significantly worse than the Motor on Block Designs.

It was suggested that these might represent two distinct groups of developmental language defects, and that recognition of the two might help to explain some of the contradictions in the literature.

I.—CLINICAL BACKGROUND.

THIS is a preliminary attempt to measure by psychological tests certain clinically observed features of children suffering from dyslalia and referred to the speech therapist between the ages of 3–6 years. It was possible for the children to be sorted into two groups on the basis of speech performance alone as already described by Greene (1960) and classified as suffering from a ' motor ' or ' language ' disorder. The speech of the youngest children at first exhibited characteristics of both motor and language disorder. This was to be expected since the normal child has the same difficulties in handling articulation and language when learning to speak, though to a lesser degree. By the age of 4 : 0–4 : 6, the pattern of a predominantly motor disorder on the one hand and a language disorder on the other had emerged. Diagnosis in each case was supported by a careful case history of general and speech development and the tests for motor and language maturation detailed below.

Motor Disorder Group.

The children classified as suffering from motor disorder had great difficulty with articulation and the accurate imitation of consonants and vowels, especially complex consonant blends such as sp, sk, str, etc. The omission and substitution of consonants followed a pattern which reduced the conventional phonetic sequences of the language to simpler forms, for example ' biscuit ' pronounced ' bitit,' ' swim ' as ' fim.' Diphthongs were frequently reduced to vowels as when ' boy ' was pronounced ' bore.' These mispronunciations were consistent as when ' sk ' was invariably substituted by ' t ' and ' sw ' by ' f.' Once the child's individual code was known, it was possible to understand what he said. Vocabulary and language development were normal for age and the Watts picture tests presented no difficulty. These children did not have difficulty with auditory discrimination and distinguished readily between right and wrong pronunciation when spoken by the therapist. Auditory memory was normal for age when tested with nonsense syllables and the standard sentences from the Terman-Merrill Intelligence Scale. Drawing was good and they had no constructional or spatial difficulty.

The developmental history and clinical testing of motor maturation supported the diagnosis of delayed motor maturation. Walking was late. The mean age of walking of the group used in the investigation (see Method following) being 16·6 months, range 14-23, whereas the mean of the language group was 14·4 months, range 10–18.

Accurate dating of commencement of speech was not possible in every case, since the mother, whose report was the sole source of information, was not able to determine when pre-linguistic babbling merged into speech ; some mothers in both groups were sure their child had never babbled in the first year.

Clinical tests of motor maturation included closing fist and thumb, thumb and finger opposition, and cutting paper with scissors (Merrill Palmer) ; hopping on one foot (Gesell) and ability to protrude and lift tongue-tip (Morley 1957), besides ability to imitate articulatory movements in consonant and vowel syllables. All children exhibited signs of late motor maturation and poor co-ordination which appeared to stem from a difficulty in isolating movements, with delay in the normal synkinetic evolution as described by de Ajuriaguerra and Stambak (1955).

The Language Disorder Group.

The language disorder group did not suffer from a similar delay in motor maturation, they crawled, sat and walked within normal limits and performed well on co-ordination tests. They had little or no difficulty in direct imitation of consonant blends and words but had considerable difficulty in *recalling* the correct phonetic sequences and also grammatical constructions in spontaneous speech. Their speech was characterised by inconsistent substitutions and omissions which rendered speech often incomprehensible even to the mother. They made unpredictable mistakes in the order of sounds within words and words within sentences. Conjugation of verbs was poor and they peformed badly on the Watts language picture tests. They had difficulty in remembering names of things and people. They could discriminate by ear perfectly when pictures of similar sounding articles were named and when the correct pronunciation was contrasted with the child's error. They could not determine whether a word was right or wrong, if presented with their own erroneous version, as in the case of the child who said ' umblella ' or ' lumbella ' for umbrella.

Auditory memory as tested by repetition of groups of nonsense syllables and the standard sentences from the Terman-Merrill Intelligence Scale was deficient. Some of the children had a constructional or spatial difficulty and failed the Merrill-Palmer tests for copying a cross, a star and the three and six-cube pyramids and assembly of the Manikin. Drawing showed an inability to appreciate form and orientation. For these reasons it was felt that these children had a deficient sense of order in time and space associated with the difficulty in retention and recall of auditory material.

II.—METHOD.

The subjects were children with severe speech defects referred for treatment to the speech therapist (M.C.L.G.). None of them had any physical malformation; clinical and audiometric examinations by the appropriate authorities had revealed no neurological or auditory abnormality. Those who fell into one of the two categories described above were referred to the Psychologist (T.G.C.), who carried out the tests. The speech therapist made the differential diagnosis between motor and language before seeing the test results, but the diagnosis was not known to the psychologist until the whole series had been tested The children had to be at least five years old when the testing was carried out, because this is the minimum age for one of the tests used (the W.I.S.C.); also they were not sent until their speech was at least intelligible to a person unfamiliar with them. All appropriate cases were referred for testing until a total of 20 was reached (8 motor (7MIF) and 12 language (8M4F)). This took nearly two years. In the testing session, the following investigations were carried out:

W.I.S.C. The I.Q. was calculated from seven subtests: Vocabulary, Similarities, Arithmetic, Digit Span, Picture Completion, Block Designs and Coding. Any child with a full I.Q. below 85 was excluded, to avoid cases where speech disability was due to inadequate general intelligence. The subtests were used for the investigation of immediate memory and ordering ability. The advantage of having standardized scores was thought to outweigh the disadvantage of using tests which were not specially devised 'pure' measures of the abilities concerned. From the way in which the scores are standardized, a child's Scaled Score on any subtests can be compared directly with his scores on other subtests, and also with the scores of other children of different ages. Digit Span was used as a measure of immediate memory, Arithmetic and Block Designs as measures of ordering ability, Arithmetic, at this level, being a measure of the child's familiarity with the sequence of numbers and the ability to apply it to simple addition, Block Designs measuring the ability to copy groups of objects in their correct spatial relationships. They were compared with Vocabulary and Similarities, which were considered the most free from 'ordering.' The other two subtests were omitted, as not falling definitely into either category. It was predicted that the Language group would be relatively poor on the Digit Span, Arithmetic and Block Designs, compared with their Vocabulary and Similarities.

Reading, etc. In Hallgren's (1950) large study of reading disability, the only factor of those studied which showed a highly significant relationship to dyslexia was the presence of speech defects. The reading of our cases was examined to see if it was poor, and to what extent. No hypothesis was made about the difference between the two groups. A mechanical reading test was used—Schonell's Graded Word Reading Test; this was thought preferable to a comprehension test, as being less dependent on intelligence. The reading age obtained from this was converted to a quotient in the usual way, so that it could be compared approximately with the I.Q. As it was not thought appropriate to test a child's reading before the age of 6, any child who was not yet 6 on the first testing was seen later, just for the reading, when he was at least 6.

Each child was asked to write his name and two simple sentences, 'a big dog' and 'I like apples,' chosen to bring out reversals or transpositions of letters (what Orton (1937) calls 'static' and 'kinetic' reversals) since these are important to theories which link speech and reading disabilities with peculiarities of laterality. This was scored simply for the presence or absence of reversals of either kind. The letters were dictated to children who could not spell words.

Laterality. The history of the study of the relationship between speech and reading difficulties and various types of laterality (left handedness, mixed or crossed laterality, etc.), is a long and contradictory one, well summarized by Vernon (1957). No prediction was made as to differences in laterality between the two groups. The assessment was made by a simple group of observations of the ' choice ' type : hand used for writing, hand for throwing (five trials), hand for catching, when instructed to use one hand (three trials), eye for counting fingers looking through a small hole in a piece of paper held in two hands (three trials, in three directions), eye used in naming pictures through a ' V-scope ' (Miles' ABC Vision Test) (five trials), and foot for kicking (three trials). Clark (1957) found kicking the best and most consistent test of footedness ; a matchbox was used instead of a ball to facilitate central placement. There are twenty items in all, and they were scored just as ' consistent ' or ' inconsistent.' The criterion for ' consistent ' was nineteen or twenty items the same side, unless the one different was the hand used for writing ; anything less than nineteen was ' inconsistent,' though as it turned out the highest score in the ' inconsistent ' group was seventeen, and the ' consistent ' all scored twenty.

The children were also asked to indicate their left foot, right ear, etc. (five trials) and to perform crossed movements (e.g., " touch your right ear with your left hand ") (three trials). These two sections were simply scored as ' correct ,' if all trials were correct, or ' incorrect.'

III.—Results.

Intelligence and Analysis of Subtests.

TABLE 1

Intelligence and Subtest Scores of the Two Groups.

	Motor (N=8)			Language (N=12)				
	Mean	S.D.	Range	Mean	S.D.	Range	t	P
Age	7·2	1·55	5·3–11·1	7·1	1·33	5·1–9·8	—	—
I.Q.	113·8	8·57	105–131	101·2	10·50	88–116	—	—
Subtests :								
Vocabulary....	12·1	1·47	10–14	11·3	2·43	8–15	0·74	>·4
Similarities ..	11·3	2·22	9–15	10·7	2·21	7–14	0·54	>·5
Digit Span	11·8	2·11	9–14	8·7	1·49	6–12	3·63	<·01
Arithmetic	13·1	2·37	9–16	9·9	1·76	7–11	3·30	<·01
Block Designs	13·3	2·28	10–18	10·6	2·50	7–15	2·30	<·05

The two groups differ significantly in the expected direction on the tests where a difference was predicted, while differences in Vocabulary and Similarities are not significant. Unfortunately, the Motor group have a considerably higher I.Q. than the Language group, so that they would be expected to be significantly better on some subtests. To eliminate the effect of I.Q., scores on each of the three critical subtests were subtracted from the mean of Vocabulary and Similarities for each individual. If the two groups are compared with

TABLE 2

COMPARISON OF SUBTEST SCORES.

	$\frac{V.+S.}{2}$ – Dig. Span				$\frac{V.+S.}{2}$ – Arithmetic.				$\frac{V.+S.}{2}$ – Blocks			
	M.	S.E.	t	P.	M.	S.E.	t	P	M.	S.E.	t	P.
Motor	−0·06	·80	0·08	>·9	−1·44	·89	1·61	>·1	−1·86	·95	1·96	>·05
Language	+2·33	·58	4·00	<·01	+1·08	·61	1·77	>·1	+0·42	·64	0·65	>·5
t (comparing the two groups)	2·47				2·42				2·06			
P	<·05				<·05				>·05 (<·1)			
	(V.=Vocabulary ; S=Similarities.)											

respect to these differences, we have an estimate of their relative efficiency on the critical tasks, as compared with the neutral ones, I.Q. being held constant. Also, from the way in which the scaled scores are derived, we know that random groups of the population would be expected to have a mean difference of nought when any two subtests are compared ; so by examining the approximation of our group mean differences to nought, we can estimate whether they differ from the general population in their relative skill on these tests, I.Q. again being kept constant. Table 2 shows the result of these two operations.

It will be seen from Table 2 that the means of the Language group all deviate from nought in the predicted direction, but only one is significant, Digit Span. In comparing the two groups all differences are in the expected direction and two are significant at the 5 per cent. level. If a one-tail test is used (as is reasonable since the results were predicted) all three are significant at the 5 per cent. level. Thus, the Language group are poorer than the Motor group on the three tests predicted, but compared with the general population on only one test, Digit Span, is their low scoring significant.

Reading. Table 3 shows the mechanical reading quotients of the two groups compared with their I.Qs. One child was lost from each group ; these both did the original testing at the age of 5, and could not be obtained for a second assessment after their sixth birthday. Neither of these could read any words, and they only named about half of the letters in the first line of the test correctly (4 and 6 respectively out of 10, I.Qs. 105 and 116, ages 5 years 4 months and 5 years 6 months), so their reading was probably retarded. The Table shows that both groups were considerably backward at mechanical reading, and both to about the same extent. Four out of seven of the Motor group and five out of eleven of the Language had reading quotients more than 30 points below the I.Q. Comparison of the (I.Q.—R.Q.) means of the two groups gives t=·50, P> ·6.

TABLE 3

MECHANICAL READING QUOTIENTS.

	Motor Group (N=7)	Language Group (N=11)
Age	M=7·44 Range=6·3—11·1	M=7·64 Range =6·1—9·8
I.Q.	M=115·00 Range=105—131	M=99·82 Range=88—115
R.Q.	M=86·14 Range=72—107	M=73·64 Range=61—87
I.Q.—R.Q.	M= +28·86 Range= +13— +38	M= +26·18 Range= +8— +45

Laterality, Knowledge of Sides and Reversals.

All eight of the Motor Group were 'inconsistent' on the laterality tests, but only four out of twelve of the Language Group. The probability of getting such an extreme division as this, by direct calculation (two-tailed) is ·008. There is no obvious pattern in the type of inconsistency shown by the Motor Group.

Putting the predominant choice for each member in the order hand, eye, foot, there are three RLR, two LLR, one RRL, one RLL and 1 LRL. Four of them also had inconsistencies within one member. In the Language Group there were eight RRR, one RRL, two RLR and one LRR. From the data given by Harris (1949) the proportion of inconsistent in the Language Group is about what would be expected in an unselected sample of children of this age group.

From Table 1, it can be seen that the Motor Group are older and more intelligent than the Language Group, so this lack of establishment of laterality cannot be ascribed to lesser general maturity.

The ability to name one's sides is probably an all-or-none ability. According to Launay (1952a), nearly all dyslexic children can do it by the age of 8 and nearly all normal children by the age of 7. Thus, it is unlikely to be discriminating after the age of 8, and we restricted this comparison to children between the ages of 5 years 0 months and 7 years 11 months. This left seven of the Motor and nine of the Language. Five Motor and only two Language were correct in simple indication of sides, and all seven were of inconsistent laterality. The nine who made mistakes were three inconsistent and six consistent. The probability of drawing all seven inconsistent from a group of ten inconsistent and six consistent (two-tailed) is ·021. Unfortunately, knowledge of sides in this age-group is clearly related to intelligence. The Mean I.Q. of the ' correct ' group is 112·3, of the ' incorrect ' group 98·8. However, there is certainly no evidence of a relationship between inconsistency of laterality and lack of knowledge of sides, and in any case the ' incorrect ' group is mostly drawn from the Language group whose I.Q. is thought to be lowered by their special disabilities. There were four 7-year-olds in the Language group, and they all made mistakes ; so they are probably worse at this than normal ; the three 7-year-olds in the motor group were all correct. Only three of the whole series were able to carry out crossed movements, two motor and one language, all three of inconsistent laterality.

In considering the question of reversals of writing, the second test of the original 5-year-olds was used. As two were missed, this gives six Motor and eight Language between the ages of 6 years 0 months and 7 years 11 months. Four of the Motor group and six of the Language showed reversals, five out of six with consistent laterality and five out of eight inconsistent. There is no relationship here between reversals and type of speech disorder or type of laterality. The degree of reversal varied greatly from just one error (' b ' for ' d '), to a complete confusion of all letters. This clearly needs a more systematic method of study.

IV.—DISCUSSION.

The prediction of the relatively poor performance of the Language group on Digit Span, Arithmetic and Block Designs was borne out in the case of Digit Span and Arithmetic. The Motor group were significantly better than the Language group on the Block Designs, but this was due to the fact that the Motor group were better on these than on other tests, not that the Language group were worse. Digit Span, the most discriminating test, is a test of 'immediate auditory memory,' but in memory tests it is always advisable to ask ' memory for what ?', and in this case it is memory for numbers *in order* ; mistakes are often mistakes of order, the correct digits being given but in the wrong sequence. So this can, perhaps, also be taken as evidence for the poor ordering ability of the Language group.

The most interesting finding is the difference between the two groups in consistency of laterality, the Motor group being all inconsistent, the Language group containing a 'normal' number of children with inconsistent laterality. This seems to provide a further definite basis for distinguishing the two types, in addition to the speech differences. It may offer a lead to the solution of the problem of the contradictory findings about laterality in reading disabilities.

We have two groups of children, who can be characterized in the following way: Group 1—Speech disability of the 'motor' type, poor reading, inconsistent laterality, no deficiency of immediate auditory memory or ordering ability; Group 2—Speech disability of the 'language' type, poor reading, 'normal' laterality, poor immediate auditory memory and poor ordering ability. In addition, knowledge of sides does not appear to be related to laterality, and tendency to make reversals in writing is not related to inconsistent laterality or to poor ordering ability as assessed by other tests; on these two points, however, the evidence is rather flimsy. In absolute terms, Group 2 are considerably worse readers than Group 1. Not one of them reaches a quotient of 90. The two groups show the same discrepancy between reading and I.Q., but it must be remembered that the I.Q. is partly composed of tests in which Group 2 are thought to do badly because of their language disability.

In spite of the negative evidence, there are so many studies showing a relationship between poor reading and peculiarities of laterality that there may be some truth in it. The easiest way of explaining the discrepancy in the findings would be to suppose the existence of two or more types of disability, and our Motor group appears to be one quite well-defined type, corresponding roughly to what Vernon (1957) calls "a class of individuals who are generally lacking in maturation" (page 109), though we should perhaps qualify it as 'motor maturation,' since they are of good intelligence, with little variation in abilities. The other group is not so well defined, the nature of their deficiency in 'ordering' ability not having been demonstrated in a completely satisfactory way. Block Designs failed to bring out clearly the clinically observed spatial difficulties of the Language Group, although they were, in fact, significantly worse on them than the Motor group. This may be due to the fact that it is a copying test. It is well known that children with severe difficulty in writing letters from memory can often copy them quite well. Perhaps a creative rather than a copying type of test would be more discriminating; de Hirsch (1957) considers a draw-a-man test to be useful for this purpose. It is hoped to investigate this further with more refined procedures.

Much of the study of developmental language defects in children has been concerned with the association of three things: (*a*) the actual disability in speech, reading and writing; (*b*) peculiarities of laterality; (*c*) defects in spatial and temporal ordering. There are *a priori* reasons for linking (*b*) and (*c*) but it seems possible that it is the attempt to do so which accounts for some of the confusion; our results suggest two groups, one with (*a*) and (*b*), one with (*a*) and (*c*), with no connection between (*b*) and (*c*). The linkage of (*a*), (*b*) and (*c*) is usually associated with the notion of a 'dominant hemisphere'; the failure of one hemisphere to become completely dominant would result directly in (*b*), and also in (*a*), since it has long been considered that speech is somehow located in the dominant hemisphere; (*c*) could be thought of as arising directly from the lack of dominance, in the way suggested by Orton, or it could be secondary, a consequence of (*b*), which might make it more difficult for the child to develop a consistent sense of direction related to his own body. However, recent findings (Humphrey and Zangwill (1952), Penfield and Roberts (1959), Zangwill (1960)) suggest that the area concerned with speech tends to be in the left hemisphere in left-handed as well as right-handed people, in other words that it is not related to dominance. There is certainly a sense in which one hemisphere is (or becomes) dominant, if only as a description of the fact that most people are more skilful with one hand than the other, and tend to choose that hand (and the corresponding foot) for skilled operations. Our Motor group certainly lack dominance in this sense, so it looks as if motor skill in speech is related to speech defect ; after all, the written symbol is based on the spoken, so if the first symbol system is faulty, the second must suffer. The Language Group, on the other hand, could be considered as having a defect in a speech or language area, not related to dominance, or perhaps in some system concerned with a more general function. Their difficulties bear at least a superficial resemblance to Gerstmann's syndrome, their poor reading and writing, poor arithmetic and sense of order, and lack of knowledge of left and right. Drew (1956) considers both Gerstmann's syndrome and ' congenital word-blindness ' to be based on a failure in Gestalt formation. In contrast, the Motor group are good at Arithmetic, and their knowledge of sides good for their age (c.f. Launay (1952a)). It is an interesting feature of this study, and worthy of further investigation, that the expectation that confusion of laterality and poor knowledge of sides would go together was not confirmed. They appear to be unrelated.

Another source of confusion in this subject is the question of reversals in reading and writing. The evidence is convincing (Vernon (1957), Chap. V) that such reversals are related to poor reading but not to laterality. If a child is going to make a mistake in writing, it is quite likely that the result will be such that it can be interpreted as a reversal of some kind. It could be that attention is drawn to this kind of error to the exclusion of other kinds. This point is clearly brought out by Miles (1961), in his discussion of the mistakes made by two children with ' developmental aphasia.' It is clear from his analysis that the disorientation which he considers to be at the root of the problem is a highly complex thing, involving space and time. It is interesting to note that his two cases of severe disability were both of consistent laterality. It is well known that nearly all children show reversals at some stage of learning to read and write (c.f., Launay, 1952). It might be that some reversals are just mistakes, while others are symptomatic of a genuine disorder of orientation.

It is recognised that much of this discussion is speculative and goes beyond the facts. However, we feel that we have distinguished two types of developmental language disorder and that investigation with the two types in mind may help to clarify some of the issues.

V.—REFERENCES.

CLARK, M. M. (1957). *Left-handedness.* London : Univ. of London Press.

DE AJURIAGUERRA, J., and STAMBAK, M. (1955). L'évolution des syncinésies chez l'enfant. *Presse Méd.*, 39, 817-819.

DE HIRSCH, K. (1957). Tests designed to discover potential reading difficulties at the six-year old level. *Am. J. Orthopsychiat*, 27, 566-576.

DREW, A. L. (1956). A neurological appraisal of familial congenital word-blindness. *Brain*, 79, 440-460.

GREENE, M. C. L. (1960). Diagnosis and treatment of late speech and language in children. *Folia Phoniat*, 12, 101-107.

HALLGREN, B. (1950). Specific dyslexia. *Acta. Psychiat. Neurolog. Supp.*, 65.

HARRIS, A. J. (1947). *Harris Tests of Lateral Dominance.* New York : Psychological Corporation.

HUMPHREY, M. E., and ZANGWILL, O. L. (1952). Dysphasia in left-handed patients with unilateral brain lesions. *J. Neurol. Neurosurg. Psychiat.*, 15, 184-193.

LAUNAY, C. (1952). Etude d'une classe d'enfants de 6 a 7 ans inaptes à la lecture. *Sem. Hôp* (Paris), 28, 1459-1463.

LAUNAY, C. (1952a). Etude d'ensemble des inaptitudes à la lecture. *Sem. Hôp.* (Paris), 28, 1463-1474.

MILES, T. R. (1961). Two cases of developmental aphasia. *J. Child Psychol. Psychiat.* 2, 48-70.

MORLEY, E. M. (1957). *The Development and Disorders of Speech in Childhood.* Chap. II, p. 17. London : Livingstone.

ORTON, S. T. (1937). *Reading, Writing and Speech Problems in Children.* London : Chapman and Hall.

PENFIELD, W., and ROBERTS, L. (1959). *Speech and Brain Mechanisms.* London : Oxford Univ. Press.

VERNON, M. D. (1957). *Backwardness in Reading.* Cambridge : Cambridge Univ. Press.

ZANGWILL, O. L. (1960). *Cerebral Dominance and its Relation to Psychological Function.* Edinburgh : Oliver and Boyd.

AN ANALYSIS OF C. W. VALENTINE'S REASONING TEST FOR HIGHER LEVELS OF INTELLIGENCE

BY H. J. HALLWORTH

(Department of Education, University of Birmingham)

SUMMARY. Answers to C. W. Valentine's reasoning test for higher levels of intelligence made by 286 men and 262 women were analysed and compared. Analysis of the item inter-correlations by the method of principal components gave a factor of general reasoning ability for both men and women. Upon rotation to the Varimax and Quartimax criteria three group factors were obtained which were comparable for the two groups of subjects. These factors are described and tentatively named. The results indicate the potential value of a detailed examination of the thinking processes involved in answering the items of this test.

I.—INTRODUCTION.

UPON the publication of his article on " The use of a new reasoning test for selection of university and training college students " (Valentine, 1961), Professor Valentine kindly made available for further analysis the test scripts of 548 university and training college students and sixth form pupils. This article reports the results of the analysis.

II.—METHOD OF ANALYSIS.

The scripts were divided into two groups, 286 completed by men, 262 by women. Using an electronic computer, two similar analyses were carried out, one for each group.

The test gives a score for each of sixteen different items. Including total test score and age, there were, therefore, scores on eighteen variables for each subject. Means, standard deviations and product-moment inter-correlations were obtained for all variables. The correlation matrix for the sixteen test items was analysed by the method of principal components, and the factor loadings were rotated to the Varimax (Kaiser, 1958) and Quartimax (Neuhaus and Wrigley, 1954) criteria.*

III.—MEANS AND STANDARD DEVIATIONS.

The mean age of the men was 22 years 3 months, and of the women 19 years 4 months. This difference was due to the fact that, at the time of testing, national service delayed the college entry of most men. The mean test score for men was 28·9, with a standard deviation of 14·5 ; the mean score for women was 24·1, with a standard deviation of 12·7. These figures are consistent with those given in the original article.

On fourteen of the test items men had a higher mean score than women. For two items women had a higher mean score, but in neither case was the difference between men's and women's means significant. It is, however, of interest that these two items are different from all others in that they each combine four characteristics :

(1) Each specifies one attribute of a logical class, using the words " None but . . . "

---

* All programmes used are for a Ferranti Mercury computer and were written by the author in mixed Autocode and machine code. (Hallworth, 1962.)

(2) Each specifies another attribute of the same logical class, using in effect the wording " All . . . "

(3) Neither of the two attributes is exclusive to the class.

(4) The answer demands an appreciation of the fact that the logical classes defined by each of the two attributes are not necessarily co-incident.

Precise information regarding adult sex differences in patterns of reasoning is not often available. This finding is recorded because, should comparable information appear from another sample using this test, or from another source, it would surely demand further investigation of the thinking processes involved.

As would be expected, the men had a higher standard deviation than the women for most items.

IV.—CORRELATIONS WITH TOTAL SCORE.

Correlations of total score with age were insignificant for both men and women, being −·05 and +·09, respectively. The negative correlation for men was probably due to the inclusion in the sample of a number of sixth formers who had recently gained university open scholarships.

The correlation of each item with the total score is given in the table below. In general the figures are very satisfactory. The lower coefficients for the last two items are due to the greater difficulty of these items, a matter which is discussed again later.

TABLE 1

CORRELATIONS OF ITEMS WITH TOTAL SCORE.

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Men	·47	·47	·40	·50	·35	·43	·41	·39	·52	·54	·53	·58	·53	·41	·32	·31
Women	·47	·41	·41	·42	·43	·27	·30	·46	·43	·43	·48	·53	·50	·35	·30	·13

V.—ANALYSIS OF THE ITEMS.

An obvious question about a test which measures reasoning ability at higher levels of intelligence is, does it measure one or more kinds of reasoning ability? The generally accepted view is that reasoning ability depends largely on 'g' (Valentine, 1961, p. 228), but that there may be differences between inductive and deductive reasoning, verbal reasoning and numerical reasoning, etc. In the original article, Professor Valentine indicates that the first four items of the test were constructed to measure inductive reasoning; the remaining items are concerned with deduction. We should, therefore, predict a factor of general reasoning ability, and a factor or factors to differentiate between the first four and the remaining items.

(a) *A General Factor.*

From the correlation matrix of the sixteen test items for each of the two groups of subjects, eight principal components were extracted. These accounted for a total of 69 per cent. of the variance for the men's group and 67 per cent. for the women's group. In each case the first component could be regarded as a general factor, accounting for 22 per cent. of the variance for the men's group and 18 per cent. for the women's group. Item loadings on this component are given in Table 2.

TABLE 2

ITEM LOADINGS ON FIRST PRINCIPAL COMPONENT.

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Men	·49	·42	·39	·50	·28	·51	·31	·30	·69	·58	·49	·70	·69	·31	·22	·20
Women	·43	·44	·49	·40	·35	·29	·15	·41	·61	·42	·40	·69	·67	·27	·18	-·06

A first component of this kind is, of course, to be expected from this method of analysis. The question is, should the analysis be accepted as it stands, or should the axes be rotated? Since there are always many possible sets of factors the decision to accept one set rather than any other must be made on psychological grounds. In the present case, acceptance of these components without rotation would give the general factor which has been postulated, and in this sense, the prediction is verified. In popular terms, this could perhaps be regarded as a typical British solution to the problem, extracting as large a common factor as possible.

However, the remaining components were not comparable across the two groups, nor did they make very good psychological sense in either case. It was, therefore, decided to rotate the axes in an attempt to discover another set of factors, more of which would have good psychological meaning.

(*b*) *Group Factors.*

For both groups, all eight components were rotated to the Varimax criterion. This is claimed by Kaiser (1958) to give results approximating to Thurstone's 'simple structure,' therefore, we may expect a more typically American solution. The results are given in Table 3. Since the method of rotation tends to spread loadings more evenly between a number of group factors, the general factor has disappeared, and there are now three factors which are fairly comparable for both groups of subjects.

Varimax Factor 1 has its heaviest loadings on items 9, 12 and 13. In order to interpret this factor psychologically, we must first note that in section two of the test there are, as explained by Professor Valentine (1961), two kinds of item: first, items requiring ability to perceive a correct argument; secondly, items requiring ability to perceive an incorrect argument *and* know *why* it is incorrect. Items 6, 9, 12 and 13 are the only items of the first type. It would probably be agreed that they can be answered more 'intuitively' than can the other times. The relatively low loading of item 6 may be accounted for by the greater complexity of argument in this item. Similarly, the relatively high loadings of items 10 and 11 may be accounted for by the brief wording of the argument and relative ease of answering. If so, the best general description of this factor is 'intuitive reasoning'.

TABLE 3

VARIMAX FACTORS FOR SIXTEEN TEST ITEMS.

	MEN				WOMEN		
FACTOR	I	IIA	IIB	III	I	II	III
Item 1	·29	·62	·15	—·12	·02	·69	·06
2	·11	·18	·76	·03	·11	·65	—·14
3	·07	·77	·05	·16	·36	·56	—·10
4	·30	·41	—·07	·10	·15	·35	·22
5	·01	·02	·14	·00	·05	·18	·13
6	·46	·08	—·22	—·07	·21	·07	·06
7	—·05	·07	·20	·10	·01	·01	—·08
8	·04	—·01	·13	·02	·16	·13	·14
9	·84	·06	—·01	·12	·77	·04	—·03
10	·53	·18	·23	·00	·20	·08	—·02
11	·29	—·08	·64	·08	·10	·23	·15
12	·66	·14	·19	·00	·74	·09	·00
13	·81	·04	·13	·03	·83	·06	·07
14	·12	—·06	.05	·11	·09	—·20	—·21
15	·11	—·12	·06	·80	·09	·11	·71
16	—·04	·27	—·01	·71	—·08	—·18	·77
Square Sum....	2·60	1·35	1·26	1·26	2·14	1·55	1·23

Varimax Factor II, for women, has heavy loadings on items 1-4, inclusive, and may be labelled 'inductive reasoning.' This factor was predicted. For men, the most comparable factor is IIA, which has, however, a low loading on item 2. Factor IIB is shown because this has a high loading on the missing item 2, also on item 11. There is no immediate logical similarity between items 2 and 11, but it may be remarked that on Factor II for the women, item 11 has the highest loading after the 'inductive' items 2-4. Here again, further investigation of the thinking processes involved might yield interesting results.

Varimax Factor III has its heaviest loadings on items 15 and 16 for both groups of subjects. These two items, the last of the test, are of a different order of difficulty from all others and yield a relatively low mean score for both men and women. Both involve relationships between five categories, and the answers involve double negation. Although Factor III was not predicted, it may be interpreted as relating to ability to deal with more complicated reasoning problems.

(c) *Further Group Factors.*

The Quartimax method of rotation has commonly been used for cognitive data on the assumption that it will bring out any common factor (Harman, 1960, p. 301). It was, therefore, used in the present instance in an attempt to obtain a general factor accompanied by other factors with psychological meaning. In Table 4 are listed the principal Quartimax factors for both groups of subjects. No general factor was obtained. Comparison with the Varimax factors shows a remarkable similarity. For both men and women, the sets of Varimax and Quartimax factors are almost precisely the same, differing by little more than the limits of precision written into the computer programmes.

TABLE 4

Quartimax Factors for Sixteen Test Items.

	Men				Women		
Factor	I	IIA	IIB	III	I	II	III
Item 1	·29	·60	·16	—·16	—·01	·68	·04
2	·12	·17	·76	·02	·12	·65	—·14
3	·11	·78	·05	·14	·37	·56	—·12
4	·33	·42	—·09	·10	·18	·35	·21
5	·04	·04	·10	·00	·06	·14	·08
6	·48	·07	—·27	—·09	·19	·05	·04
7	—·02	·08	·15	·09	·00	·00	—·08
8	·04	—·10	·10	·02	·15	·15	·18
9	·84	·05	—·02	·11	·77	·02	—·04
10	·53	·17	·23	—·03	·21	·08	·00
11	·31	—·09	·61	·08	·10	·26	·17
12	·67	·12	·17	—·02	·73	·08	—·02
13	·82	·03	·12	·02	·84	·05	·05
14	·10	—·05	·04	·07	·09	—·20	—·21
15	·14	—·09	·06	·80	·10	·11	·68
16	—·04	·30	·00	·71	—·07	·16	·79
Square Sum....	2·69	1·34	1·20	1·24	2·15	1·53	1·29

VI.—Conclusion.

As is usual with factorial analyses, there is more than one set of factors which have psychological meaning. Accepting the principal components as they stand, there is a general factor of reasoning ability accounting for approximately one-fifth of all the variance. Rotating to either the Varimax or Quartimax criterion, there are three factors which are related to features of the test items and to the reasoning abilities which they apparently demand. These factors, and certain differences between scores for men and for women, suggest that it would be valuable to have a detailed examination, possibly along the lines used by Duncker (1945) and Peel (1962), of the thinking processes involved in answering the items of this test.

VII.—REFERENCES.

DUNCKER, K. (1945). On problem-solving. *Psychol. Monogr.*, 58, No. 270.

HALLWORTH, H. J. (1962). Computer programmes for psychologists. *Bull. Brit. Psychol., Soc.*, 48, 16-21.

HARMAN, H. H. (1960). *Modern Factor Analysis*. Univ. Chicago Press.

KAISER, H. F. (1958). The Varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187-200.

NEUHAUS, J. O., and WRIGLEY, C. (1954). The Quartimax method. An analytic approach to orthogonal simple structure. *Brit. J. Stats. Psychol.*, 7, 81-91.

PEEL, E. A. (1962). Learning and meaning. *Bull. Brit. Psychol., Soc.*, 48, 1-9.

VALENTINE, C. W. (1954). Reasoning tests for higher levels of intelligence. Edinburgh: Oliver and Boyd.

VALENTINE, C. W. (1961). The use of a new reasoning test for selection of university and training college students. *Brit. J. Educ. Psychol.*, 31, 227-231.

THE DEVELOPMENT OF CHILDREN'S IDEAS ABOUT COUNTRY AND NATIONALITY

PART I: THE CONCEPTUAL FRAMEWORK.

BY GUSTAV JAHODA
(*University of Glasgow*)

SUMMARY. A random sample of children aged 6–11 in four Glasgow schools was interviewed and tested, two situated in middle-class and two in working-class districts. The formation of concepts relating the series of units Glasgow–Scotland–Britain is traced in terms of both verbal formulation and spatial representation. The difficulty for younger children of such concepts as 'town' and 'country' is demonstrated. The emergence of national identification is described, indicating the major sources of initial confusion.

I.—INTRODUCTION.

THIS study was originally inspired by Piaget (1928, 1951), who attempted to establish developmental stages for children's concepts of country and nationality. Unlike most other replications concerned with more clear-cut spheres of thought, such as the work of Lovell (1961) on mathematical and scientific concepts, the findings were in several fundamental aspects at variance with those of Piaget. The theoretical problems raised thereby will be dealt with separately. Here the aim is a two-fold one: to analyse the gradual emergence of concepts and relationships, and to offer tentative developmental norms for children of contrasting socio-economic background.

II.—METHODS.

(1) *The Sample.* The first stage of sampling was a 'purposive' one: two primary schools were selected from well-to-do middle class areas, where parental occupations are mainly clerical, professional and administrative; in contrast, two other schools were located in a high-density and now much dilapidated working class district, where semi- and unskilled occupations predominate.

Within the schools, a sampling frame was constructed from the registers, pupils being divided according to age and sex irrespective of their form. The minimum age was fixed at six, because it had been obvious from preliminary trials that it would have been difficult to administer the somewhat lengthy procedure to the youngest children; and it was felt that uniformity was essential if valid comparisons were to be made. After excluding any children not of Scottish parentage, selection was done at each age-level by means of a table of random numbers. three boys and three girls always forming the unit. This meant that the sample was widely scattered, only a small number being taken from any one form. The object of this was to minimize the possibility of communication among the participants, and there is every indication that this was satisfactorily achieved. The general plan is, perhaps, easier to follow when set out schematically:

Middle Class District (*M*)				Working Class District (*W*)			
School A		School B		School C		School D	
Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls

(Ages: 6– 7– 8– 9– 10– 11–)

There were thus :

2 (social classes) × 2 (schools) × 2 (sexes) × 6 (age levels) × 3 (individuals) = 144.

Unless otherwise specified, all subsequent tables will be based on this total of 144 children. Although sex had to be one of the controlled variables, significant sex differences failed to emerge. Hence results will not be shown separately for boys and girls. This does not necessarily mean that there are no contrasts between the sexes in this field ; in fact, some of the differences were at least suggestive. However, in view of the relatively small numbers, dictated by the time-consuming nature of the interviews, it was felt that one would not be justified in putting much stresss on this aspect.

(2) *Interviewing and Testing*. The children were seen individually. Each session lasted between 35 and 70 minutes, the mean being about 45 minutes. In order to arouse and maintain the interest even of the younger children, the procedure was devised so as to present considerable variety.

In an investigation of this kind, the handling of the situation is of paramount importance. The procedure was not started until the children had been put at ease by some preliminary chatting, and, in fact, an atmosphere of informality was aimed at throughout. The schedule itself was indended as a guide, not a straight-jacket. The questions in the schedule were certainly always used in their exact wording ; but the possibility of purely verbal misunderstandings was constantly kept in mind. Hence, particular questions were often re-phrased and put again, taking care at the same time not to introduce any suggestions. The purpose was always that of exploring the degree of clarity and the range of various concepts for the child.

Naturally, this is a difficult task, and some influence of the observer upon what is being observed is unavoidable. As has been rightly said,* the very attempt to ascertain the child's capacity for comprehension is itself a learning situation. When a child is faced with a pointed question concerning some relationship for the first time, this may be the very stimulus which suddenly enables them to perceive how the elements fit together, so that a correct answer is produced after reflection. This did occasionally happen, but any error arising thereby will, from the present stand-point, be in the conservative direction. In other words, it will lead to an *overestimate* of the capacity to grasp a particular concept at any given age level.

III.—Developmental framework.

(i) *Geographical Stages.*

There is general agreement that a child's intellectual grasp of his environment begins in his immediate vicinity and only gradually extends outwards. One can think of this as a series of concentric circles, with the child at the centre. Such a model, admittedly far too simple and neat to be a realistic portrayal of the child's perspective, may, nevertheless, serve to define the first task : it is that of determining the extent of each child's effective range. The word 'effective' means in this context the range within which the child is capable of correctly apprehending relationships.

One of the chief measures to be introduced here is what with some diffidence may be called the child's geographical stage, or 'G-stage' for short. This is derived from the verbally formulated relationships, in response to questioning.

---

* Lovell and Ogilvie (1960).

The measure was arrived at after attempting, without successs, to apply Piaget's scheme. The one adopted here involves the use of what might be called ' natural ' units for the purpose of categorization. Whilst not eliminating the element of arbitrariness entailed in the slicing up and pigeon-holing of what is essentially a continuum, it does at least tend to reduce it. There were, of course, a number of marginal cases, but the large majority could be easily and unambiguously classified. The categories, or G-stages, were as follows :

G.I. : No conception of Glasgow as a unitary whole.
G.II : Conception of Glasgow as unitary whole, but no conception of it as part of Scotland.
G.III : Conception of Glasgow as part of Scotland, but no conception of Scotland as part of Britain.
G.IV : Glasgow-Scotland-Britain relationship correctly expressed.

The ideas of children at these four stages will now be illustrated with examples. The questions by the interviewer are shown in brackets ; the symbols after each quotation stand for type of school and age.

G.I.

It was characteristic of children at that stage that they seemed to think of Glasgow as some kind of vague entity near-by, though not usually including their own immediate geographical location.

(*Where is Glasgow*?)
Beside the playground, in X street. (*W*6)
Its up by the park there—you go round the corner. (*W*6)
Down in Y avenue. (*M*7)
I don't know the way to Glasgow. (*W*7)

Although nearly all were familiar with the word ' Scotland,' their notions of this concept were apt to be nebulous in the extreme. Several main tendencies are distinguishable : (*a*) the fusion of Scotland with Glasgow ; (*b*) the belief that Scotland is somewhere *outside* Glasgow ; (*c*) random associations called up by the word ; and also combinations of these.

(*What is Scotland*? *Where is Scotland*?)
I get muddled up—is that not Glasgow ? Yes, it's just Glasgow. (*W*6)
My granny stays in Scotland. (*Do you go and see her*?) Yes, I go up at 3 o'clock every day. (*W*7)
I was born in it—it's the name of the town. It's just outside Glasgow. (*W*8)
Scotland is the capital of Edinburgh. It's in Glasgow. (*M*7)
Scotland is a street. (*Yes, Scotland Street, but have you heard of a place called Scotland*?) Yes, Scotland the Brave, it's up in the Highlands. (*W*7)
The Kiltie-Dancers—on the telly—I saw it last night. (*Where is Scotland*?) Away far. (*W*6)

The majority could say nothing about Britain at all, beyond perhaps the fact that they thought they had heard of it. If it was conceived as a place at all, it was clearly a very remote one.

(*What is Britain*? *Have you heard of Britain*?)
I've heard the word, but I don't know what it's like. (*W*6)
Yes. (*What do you think it is*?) Our next-door neighbour went to it and brought some dolls home. (*M*7)

It was there that the men fought the war. (*W*8)

In general, it may be said that at this first stage the organization of geographical concepts is minimal, responses being mainly in the nature of diffuse and undirected associations.

G.II.

The chief difference between this stage and the preceding one is that the children's answers showed an awareness of being within the unit called Glasgow, which was no longer thought to be located somewhere else. Responses to the question "Where is Glasgow?" were fairly stereotyped, usually taking the form "It's here, where we are." There were, moreover, a few replies which may be regarded as transitional to the subsequent stage, though they undoubtedly belong to G.II because of their inherent contradictions.

Glasgow is in Scotland. (*Where is Scotland?*)
Scotland is a country outside Glasgow. (*W*7)
Glasgow is in the middle of Scotland. (*What is Scotland?*) It's a place where the people play the bagpipes. (*Where?*) Away in the Highlands. (*M*7)

The remainders' ideas about Scotland were largely similar to those of G.I, including the notions that Scotland is 'in England' or 'part of America' (both *M*6). A higher proportion ventured some statement about Britain, but the level of accuracy was usually not much higher.

(*What is Britain?*)
It's a city in England. (*W*7)
It's a big town: mountains are there, and they climb on it. (*W*7)
I don't think I've been there. (*M*6)
It's a city in Scotland. (*M*9)

The major advance, therefore, is confined to the consolidation of the concept of Glasgow as an entity in which one lives and which constitutes one's surroundings.

G.III.

Here the stock answer became "Glasgow is in Scotland." or words to that effect. Scrutiny of the total response patterns suggests, however, that one must be careful not to read too much into this. It is true that replies did not go astray until after the question "Where is Scotland?", but as will shortly become apparent the foundation probably tends to be somewhat shaky.

(*What is Scotland? Where is Scotland? What is Britain?*)
Scotland is a town made up of countries and towns. It's up north. Britain is a place where you live in, a big, big, piece of land surrounded by water. (*M*7)
Scotland is a country. It's in the world. Britain is really the whole continent. (*M*8)
Scotland is a country. It's beside Wales and Ireland. Britain is mostly called Great Britain; it's sort of Scotland—another name for it. (*M*8)
Scotland is a country. Britain is the United State. (*What do you mean?* Long pause.) Of America. (*W*9)
Scotland is a country. (*Where?*) Can't say. Britain is a lot of different countries. (*What countries?*) Glasgow, London, France. (*W*11)

When one considers these and similar answers the phrases " Glasgow is in Scotland " and " Scotland is a country " begin to have a suspicious air of an empty formula. Additional evidence shows that this is not always justified, yet in the main one seems to be faced with the parrot recital of terms learnt without adequate understanding. In other words, grasp of relationships has failed to keep pace with the acquisition of a new vocabulary. It is thus, perhaps, doubtful whether G.III should really always be accepted as a genuine stage regarding concept formation ; it might be safer to look on it as a temporary transition between G.II and G.IV.

G.IV.

This can be dealt with very briefly. In order to qualify, a child had to convey some understanding of Britain as well-defined unit. The first of the quotations below represents roughly the acceptable minimum.

> Glasgow is in Scotland. Scotland is a country in Great Britain. Britain is some countries joined together. (*M*6)
>
> Glasgow is in the Central Lowlands of Scotland. Scotland is a country. It's just north of England. Britain is Scotland, England, Wales and Northern Ireland. (*W*11)

With G.IV children have achieved the normal adult standard in Britain.

The levels reached by various sub-groups are summarized in Table 1. In this and most subsequent tables results are presented separately for three age-ranges, comprising twenty-four children each ; within age-ranges there is a further sub-division into *M* and *W* schools. This produces a general picture of developmental trends as influenced by age and socio-economic background.

TABLE 1

GEOGRAPHICAL STAGES REACHED BY TYPE OF SCHOOL AND AGE.

Geographical Stages	*M* Schools			*W* Schools		
	6–7	8–9	10–11	6–7	8–9	10–11
G.I.	4	1	—	16	5	—
G.II	13	3	—	8	10	3
G.III	4	7	2	—	4	12
G.IV	3	13	22	—	5	9

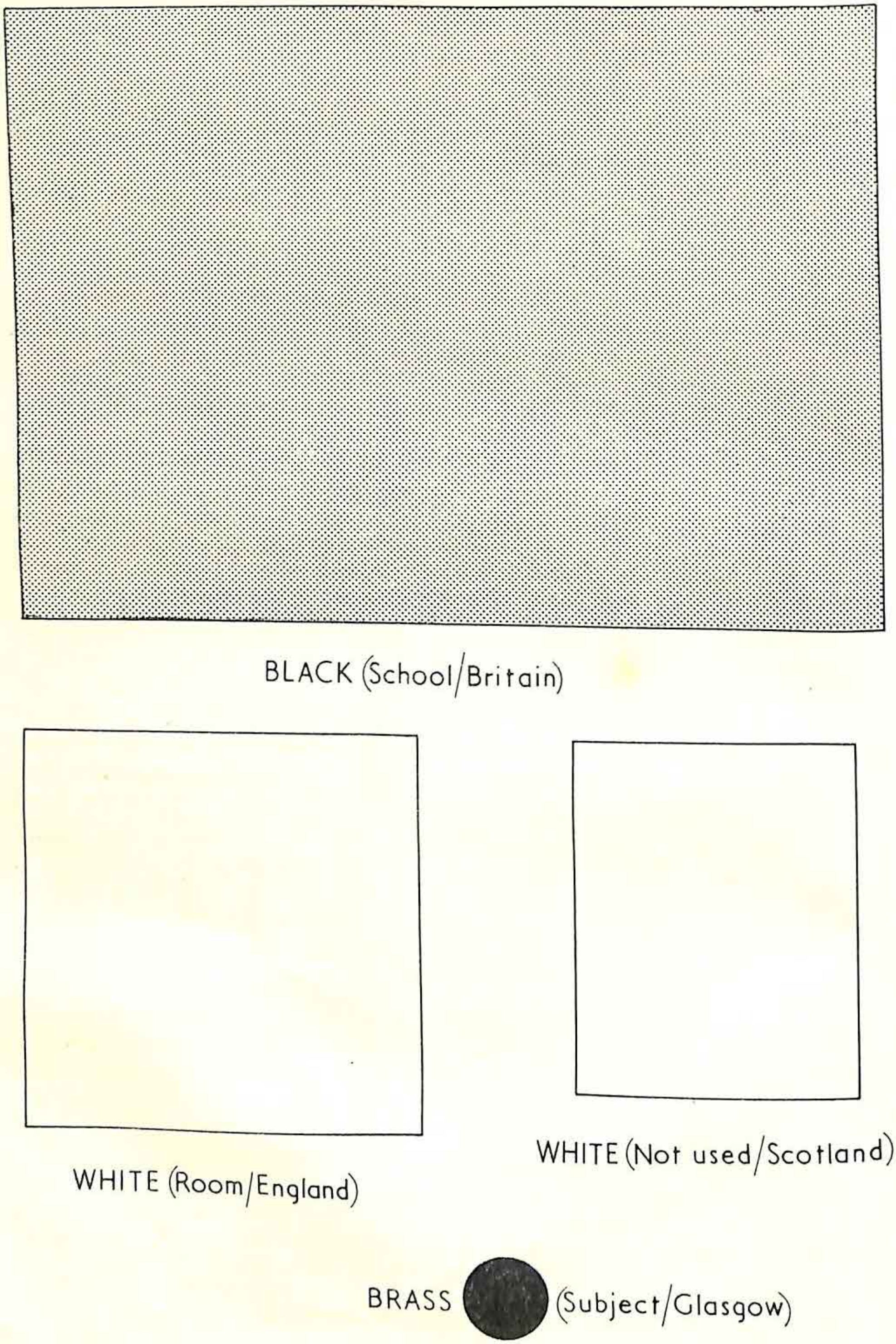
There is a highly significant difference (P less than ·001) between the distribution of stages in *M* and *W* schools. It may also be noted that G.III contains a smaller total than the stages preceding and following it, thus affording some support to the view that it might well be a relatively brief transition stage.

(ii) *Spatial Representation of Geographical Relationships.*

The procedure used in determining G-stages is open to the objection that it is conducted entirely at the verbal level, and in fact, some doubts have been expressed on this score as regards G.III. A special test dealing with the appropriate spatial concepts was, therefore, devised. It was administered in two parts the first being a preliminary training session designed to familiarize the children in a simple manner with the symbolic relations involved.

FIG. 1

MATERIALS USED IN THE SPACE TEST.



The material employed, seen in Fig. 1, consisted of one black and two white plastic boards, and a brass disc. For the try-out the smaller of the white pieces was left out, and the child asked to reproduce the immediate situation, i.e., self-in-room-in-School ; this, is of course, analogous to Glasgow-in-Scotland-in-Britain. Any child failing in this preliminary task was given a detailed explanation and demonstration with the aid of the materials. This was necessary with 12·5 per cent. of all *M* children and 37·5 per cent. of *W* children (P less than ·001).

The decisive test came later, immediately following a musical test, so as to minimize carry-over from the questioning. This time the child was given all four pieces, with the information that the largest one was supposed to be Britain. The task was (*a*) to decide which of the others represented Glasgow, Scotland, and England, respectively, and (*b*) to assemble them correctly. The criterion for failure under (*a*) was simply not choosing the small brass disc for Glasgow ; only a child making this mistake, whatever the subsequent performance, was classed as S.I.

In order to qualify for success in positioning, three conditions had to be satisfied : brass disc on top of one of the white pieces ; white pieces side by side, not on top of one another ; white pieces within the area of the black one. Anything not conforming to this was again classed as failure, S.I. The most common failures are illustrated in Fig. 2. Among the youngest children, side-by-side placement of all the pieces often happened (S.I.—A) ; overlapping sometimes occurred (S.I.—B), or simple building up (S.I.—C), which was an attempt to reproduce the earlier practice trial whilst coping with the additional piece. The most frequent failure of all was something like S.I.—D, namely, the brass disc flanked by the two white pieces ; from the conceptual point of view, this suggests that such children thought of Glasgow and Scotland as two separate units, without as yet any notion of spatial inclusion.

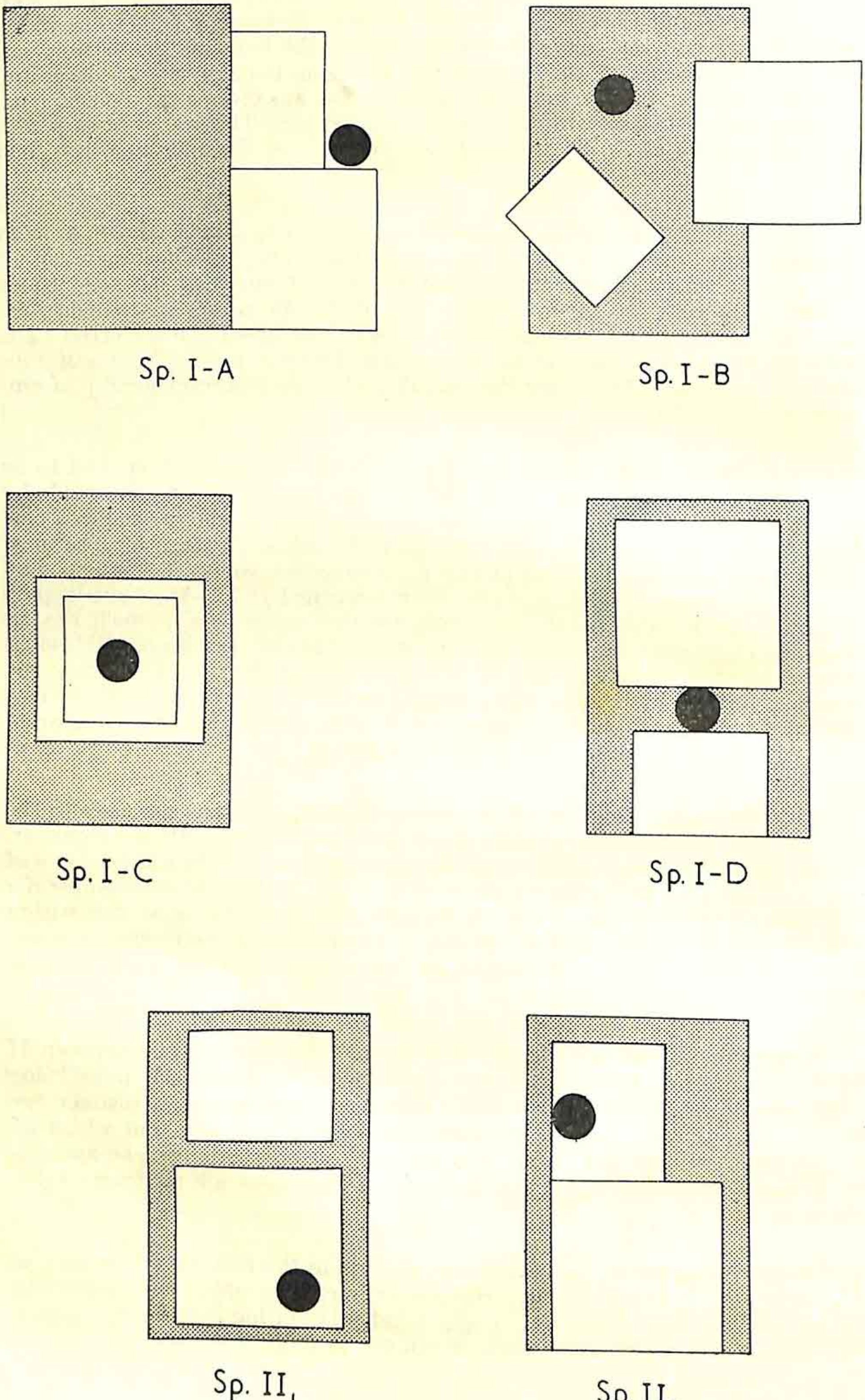
All children fulfilling the above conditions were scored as successes (S.II), ignoring both identification of white pieces and compass orientation. However, a further minor distinction was made between children who had at least one of these wrong (S.II$_1$), and those who were right in both (S.II$_2$). In the example of a successful solution shown in Fig. 2, the location of the Glasgow disc within Scotland was obviously carefully chosen ; but the exact positioning was not considered in the scoring, as it would have been difficult to know where to draw the line.

Results are set out in Table 2, and it appears that the contrast between *M* and *W* is as sharp as it was in the case of G-stages, the overall difference being highly significant (P less than ·001). Closer scrutiny reveals a distinct age pattern : at 6–7 there is little to choose between *M* and *W* ; but whilst *M* children progressed steadily from then onwards, *W* ones exhibited no appreciable advance until after the age of 9, and then they were still far from having caught up.

Another observation may be made concerning the relative frequencies of S.II$_1$ and S.II$_2$: it looks as though the former were less stable. In other words, once the basic nature of the Glasgow-Scotland relationship is grasped, compass orientation and size soon come to be sorted out as well.

FIG. 2

TYPES OF FAILURE AND SUCCESS IN THE SPACE TEST.



Sp. I-A

Sp. I-B

Sp. I-C

Sp. I-D

Sp. II_1

Sp. II_2

TABLE 2

SPATIAL STAGE REACHED BY TYPE OF SCHOOL AND AGE.

Spatial Stages	*M* Schools			*W* Schools		
	6–7	8–9	10–11	6–7	8–9	10–11
S.I	20	8	4	22	21	12
S.II$_1$	1	6	4	2	2	3
S.II$_2$	3	10	16	—	1	9

(iii) *Comparison of the Two Approaches.*

This is provided in Table 3, which shows a regular and orderly progression. The relationship is most clearly apparent from the bottom row, which gives the percentages of complete failure on the space test for each G.-stage. The first sizeable reduction in the failure rate occurs at G.III, and there is a spectacular drop between G.III and G.IV.

TABLE 3

RELATIONSHIP BETWEEN GEOGRAPHICAL AND SPATIAL STAGES.

S-Stages	G-Stages				Totals
	G.I	G.II	G.III	G.IV	
S.I	25	35	21	6	87
S.II$_1$	1	1	5	11	18
S.II$_2$	—	1	3	35	39
Totals	26	37	29	52	—
Per cent. S.I. for each G-stage	96%	95%	72%	12%	—

The comparison offers a satisfactory confirmation of the interpretation already put forward. Stage G.III is clearly one of transition, in the course of which the verbal statement " Glasgow is in Scotland " remains largely an empty formula, devoid of an adequate conceptual appreciation of the part-whole relationship involved. On the other hand, a large majority of those able to given expression to the broader Glasgow-Scotland-Britain pattern have, in fact, achieved such a grasp.

Before leaving this important section, attention should also be drawn to the wide variations in ages at which a particular stage is attained, which emerges from Table 1 and 2. It is noteworthy, moreover, that only half the *W* children of 10 and 11 have reached an adequate concept of spatial relations, a fact which tends to be concealed by the effects of rote learning.

(iv) *The Concept of ' Large Town.'*

It may have been observed in connexion with some of the quotations that children were apt to describe countries such as Scotland and England as 'towns.' The converse error was also made occasionally, though there were fewer opportunities for this; a confusion of the kind is certainly implicit in such statements as "Scotland is the capital of Edinburgh."

Immediately following the topic of Glasgow, children were asked to name some other 'really big towns.' This was an attempt to discover the range of notions conveyed by the phrase at different ages. In evaluating the findings, summarized in Table 4, one limitation of this approach must be kept in mind: the total number of responses given by each child was not controlled, and this is a possible source of variation. However, as the general pattern seems to make good sense, it cannot have been of great importance.

Among the youngest children, nearly one-third were unable to given any answer, or mentioned street names; the remainder of this group produced a large and ill-assorted variety of responses, only two of them entirely correct. In the intermediate age-group, most of the crudest errors had dropped out, leaving mainly a confusion between small towns, large towns and countries. Finally, at the ages 10-11, the chief remaining mix-up was between small and large towns. As before, *M* children were generally greatly in advance over *W* ones (P less than ·001).

TABLE 4

TYPES OF RESPONSES FOLLOWING REQUEST TO NAME 'REALLY BIG TOWNS.'

	M Schools			*W* Schools			Rounded Percentages
	6–7	8–9	10–11	6–7	8–9	10–11	
Large towns only*	2	4	11	—	2	8	19%
Large and small towns	5	9	10	—	3	3	21%
Countries, large towns, and small towns	8	7	3	5	8	9	28%
Countries, small towns and streets	5	4	—	8	8	3	24%
Streets only	—	—	—	6	—	—	
Unable to name anything	4	—	—	5	3	1	9%

On the basis of these findings, one may hazard some inferences about elements of the thought processes involved. With the least mature group, the question evokes just any kind of geographical names that happen to be familiar to the child; in other words, the concept of large towns is so diffuse that it is taken to cover a variety of units, from streets to countries. If these concepts have not yet become differentiated, one can hardly expect the child to establish any kind of relationship between them, as was the task with regard to Glasgow, Scotland and Britain.

Gradually some boundaries begin to take shape, and the child moves into a zone of transition. Whilst the difference between towns and countries is not fully understood, some children are able to produce at least an appropriate verbal formula to designate the relationship (e.g., "Glasgow is in Scotland") others, though fewer, are beginning to apprehend the spatial relations.

---

* A 'large' town was defined as having a minimum population of 150,000.

The final stage, in terms of the classification adopted here, is the achievement of a sharp discrimination of units. This permits both an insightful verbal formulation, as opposed to parrot recital, and correct abstract representation of spatial arrangements.

If such an interpretation is valid, it follows that the notion of what constitutes a ' really big town ' ought to be closely associated with both G and S stages. Table 5 indicates that this is, in fact, the case (P less than ·001).

TABLE 5

TYPES OF ' REALLY BIG TOWNS ' RESPONSES IN RELATION TO G AND S STAGES.

	Response Categories		
	Large towns only + small and large towns	Small towns and/or large towns and countries	All other responses
Numbers in each category	57	46	41
Per cent. in G.III and G.IV	81%	54%	24%
Per cent. in S.II	67%	37%	5%

IV.—COUNTRY AND NATIONALITY.

(i) *Spontaneous National Identification.*

The first question asked when the interview proper started was " What are you ? " The object was to elicit the dominant manner in which children tended to view themselves, before anything else could introduce a bias ; in particular to find out to what extent, if any, nationality entered into what are commonly called children's ' self-images.'

Results can be put very briefly : some nine-tenths confined themselves to their sex-role, i.e., they said " I am a boy/girl." Further probing brought out mainly qualifications of the statement already made (e.g., " I am a schoolboy," " I am a good girl "), or with older and/or brighter children, some superordinate categorization like ' person ' or ' human being.' Out of the total sample only just over 3 per cent. described themselves as ' Scottish,' all being from *M* schools. It may also be mentioned that the same proportion referred to their religion.

This may be compared with the findings of a study by Lambert and Klineberg (1959) carried out about the same time, and using the identical question. Three sets of English children (among other nationalities) aged 6, 10 and 14, respectively, were included in the survey. Omitting the oldest, the combined proportion of the others who called themselves English came to about 3 per cent., almost exactly the same. This is, perhaps, somewhat surprising, as one might have anticipated that Scottish children would show a relatively greater awareness of their nationality. It is possible that such a difference would appear later, as the 14-year-old English children showed a great upsurge in the frequency of giving their nationality.

(ii) *General Inquiry about Country and Nationality.*

For most adults the phrase ' my country ' has powerful emotional connotations. The problem of the growth of such sentiments is merely touched upon in this investigation, which focusses primarily on cognitive development. It was concerned, however, with the emergence of understanding of the object to which such sentiments refer.

Before Scotland or Britain had been mentioned by the investigator, children were asked " What is your country ? " It is apparent from Table 6 that the youngest children often had no idea, some naming either their street or local district. A frequent answer was ' Glasgow,' and this may have been influenced by the preceding talk about the home town. With rising age, ' Scotland ' gained increasing prominence, whilst the small proportion who named Britain remained fairly constant. Differences between *M* and *W* schools were significant (P less than ·01). The general pattern fits in well with the view that children begin with an awareness of their immediate locality, which becomes progressively extended outwards, *pari passu* with a sharperned discrimination concerning such geographical terms as ' town ' and ' country '

TABLE 6

DISTRIBUTION OF ANSWERS TO THE QUESTION ' WHAT IS YOUR COUNTRY ? ' BY TYPE OF SCHOOL AND AGE.

	M Schools			*W* Schools		
	6–7	8–9	10–11	6–7	8–9	10–11
Don't know or miscell.	4	2	—	12	7	3
Glasgow	7	2	—	3	4	4
Scotland	11	19	22	8	11	15
Britain	2	1	2	1	2	2

The question " What is your nationality ? " held few surprises. Among *M* children only one in the 6-7 group understood it, as compared with one-third of those aged 8-9 and three-quarters of the oldest ones ; corresponding figures for *W* children were, respectively, nil, a single one and one-third. Most of those who understood the question gave a correct reply, i.e., ' Scottish ' or ' British,' but a few said ' England ' or ' English.'* The probable reason for such answers will be encountered shortly.

(iii) *Understanding of Relationships.*

After reference had been made to Scotland and Britain, children were asked specifically whether they were Scottish, and whether they were British. As is evident from Table 7, a majority of all ages gave an affirmative reply to at least one of these questions. Moreover, there is no significant difference between *M* and *W*, indicating that a verbal knowledge of nationality is acquired at roughly the same rate in both kinds of environment. Most of those aged 10-11 acknowledged that they were both Scottish and British ; the problem is, however, to what extent this corresponded to a grasp of the implicit part-whole relationship.

---

* It will be remembered that children not of Scottish parentage had been excluded from the sample.

TABLE 7

DISTRIBUTION OF COMBINED ANSWERS TO THE QUESTIONS 'ARE YOU SCOTTISH? ARE YOU BRITISH?' BY TYPE OF SCHOOL AND AGE.

	M Schools			*W* Schools		
	6–7	8–9	10–11	6–7	8–9	10–11
'No' or 'Don't know' to both	6	2	—	9	7	—
Scottish 'Yes' and British 'No' or 'Don't know'	8	4	—	5	4	4
British 'Yes' and Scottish 'No' or 'Don't know'	2	1	—	3	3	1
'Yes' to both	8	17	24	7	10	19

In order to test this, a follow-up was used with children who had said they were both, and with those who were positive about one but uncertain about the other. All these were asked about the possibility of being Scottish and British at the same time, and invited to explain their reasons.

The outcome, summarized in Table 8, is instructive. The most frequent reason offered for rejecting the possibility was mutual exclusiveness—" you can't be two different things at once." The other prominent feature, used as a justification on either side, was language. In fact, language and nationality were commonly mixed up, which probably led some of the children to say 'English' when asked about their nationality, as mentioned earlier. A few quotations will illustrate the prevalent state of confusion, the first pair below having answered 'Yes' and the second 'No':

> British is your language and Scottish is your accent. (*M*8)
>
> Down in England you speak British and here in Scotland you still talk British. (*W*10)
>
> I speak English—I'm really Scottish. (*M*6)
>
> You can't talk Scottish and British at the same time. (*W*9)

The first two happen to be consistent, though obviously without any glimmer of understanding; the last two, having declared previously that they were Scottish and British, had changed their position, yet there was no indication that they were aware of it. On the other hand, a small number of children did show such an awareness. For instance, one decided " You can't talk different ways," and reverted to Scottish only; another, possibly with a notion of Gaelic at the back of his mind, came to the conclusion that he was British but not Scottish because " All the Scottish speak a different language to us"; one altered his answer from 'No' to 'Yes,' explaining " Well—you can speak all the different languages." Lastly, an 11-year-old, after an initial emphatic 'No,' hesitated and finally reached the correct answer. This is an example of the interview as a learning situation, in the course of which new insights may be gained.

TABLE 8

IDEAS ABOUT THE RELATIONSHIP BETWEEN 'SCOTTISH' AND 'BRITISH.'

Can one be both ?	Explanation given	N
Don't know	None	4
No	None	7
	'Can't be two things at same time'	16
	Different language	6
	Other	3
Yes	None	4
	Same language	7
	Other	3
	Correct explanation	44

The miscellaneous answers contained some curious specimens. Some thought that Scottish and British were just different names for the same country ; one explained that his father was Scottish and his mother British, and a 6-year-old hit upon an ingenious solution : " One week I'm Scottish and the next week I'm British."

The lesson of this follow-up is that replies to the nationality questions cannot be taken at their face value. When challenged, about a third of the children said that one could not be British and Scottish at the same time, thereby mostly contradicting their earlier responses and showing that these had not been rooted in a genuine understanding ; others continued to affirm the possibility of being both, yet without being able to provide an adequate justification ; and less than half were able to demonstrate their grasp of the underlying relationship.

V.—REFERENCES.

LAMBERT, W. E., and KLINEBERG, O. (1959). A pilot study of the origins and development of national stereotypes. *Internat. Soc. Sci. J.*, 11, 221-238.

LOVELL, K. (1961). *The Growth of Basic Mathematical and Scientific Concepts in Children.* London : University of London Press.

LOVELL, K., and OGILVIE, E. (1960). A study of the conservation of substance in the junior school child. *Brit. J. Educ. Psychol.*, 30, 109-118.

PIAGET, J. (1928). *Judgment and Reasoning in the Child.* London : Kegan Paul.

PIAGET, J., and WEIL A. (1951). The development in children of the idea of the homeland and of relations with other countries. *Internat. Soc. Sci. Bull.*, 3, 561-578.

(*Manuscript received,*

PSYCHO-LOGIC: A POSSIBLE ALTERNATIVE TO PIAGET'S FORMULATION

By G. HARRY McLAUGHLIN
(Northampton College of Advanced Technology, London)

Summary. The limitations of children's reasoning observed by Piaget can be described by a restricted logical system, called a psycho-logic. It is shown that psycho-logic can be simply represented by a small set of finite Boolean algebras. These are interpreted to suggest that limitations of reasoning correspond to limitations of immediate memory or attention.

Piaget (1947) is particularly noted for the discovery that children behave as if their thinking conformed to the rules of a psycho-logic, meaning a logical system too restricted to permit certain valid deductions to be made. He also observed that a child's ability to reason progresses through an invariable sequence of four main stages, outlined below. He accounts for this sequence by suggesting that the development of higher stages of reasoning depends upon the child achieving concepts which are more and more abstract.

Nearly all the generalisations on the development of reasoning given by Piaget (1937), Piaget and Inhelder (1941, 1959) and Inhelder and Piaget (1955), have been supported by numerous careful investigations, such as those frequently reported in this *Journal* by Peel and his co-workers.

To describe the four stages of intellectual development, Piaget formulates four levels of psycho-logic. Each level is a logic so restricted that it yields only those inferences which can be made by a subject at the corresponding stage. Furthermore, Piaget advances the hypothesis that if, at any level, a formula can be derived from another formula in psycho-logic, a subject at the end of the appropriate stage can carry out cognitive processes corresponding to both formulas.

Psycho-logic has been set out in detail by Piaget (1949); particular aspects of it have been exhaustively treated by Piaget (1942, 1952); and it is summarised by Piaget (1947, 1953). The last-mentioned volume has an introduction by Mays explaining with admirable clarity the logical terms used in psycho-logic. Unfortunately, as Beth (1950) noted, Piaget's own expositions of his system contain serious confusions and are consequently hard to follow. Grize (1960) has expressed the system in a rigorous form, but this is even more formidable to the non-logician, and few of its terms appear to have psychological referents.

The difficulty of Piaget's system seems to stem from his attempt to translate directly into logical terms the psychological notion of increasing conceptual abstractness. This gives the system a structure different from that of conventional logics.

However, it will be shown that a psycho-logic can be adequately formulated in terms of the algebra of classes, which is explained in most introductions to modern logic (e.g., Langer 1953). It will also be shown that each level can be simply defined by the number of different classes that are distinguished simultaneously. Furthermore, this number can be directly identified with a clearly measurable psychological characteristic—memory span.

Briefly, the argument is that a subject solving an intellectual problem must retain a number of items at the same time, the number depending on the logical complexity of the problem ; and that the number of items a child can retain simultaneously increases as he grows older ; therefore, as the child matures, he can solve more complex problems. In other words, each of the four stages in the development of thought is characterised in part by the number of items that a child can retain at the same time. It should be noted that increases in retentivity are claimed to lead to changes only in the logical aspect of thinking : no doubt changes in the quality of thinking in its other aspects involve further factors. Furthermore, an ability to retain a certain number of items is claimed only to be a necessary condition for making certain kinds of inference : it is not proposed as a sufficient condition.

There are at least two ways in which the number of items retained simultaneously could increase with age. The older child might either be able to remember more items after a single presentation, or be able to attend to more items while they are being presented. Even more probably, memory capacity and attention capacity increase together, perhaps because they are two aspects of the same process. For convenience of exposition, a child's growing ability to consider a certain number of items simultaneously will usually be treated simply as a function of enlarging immediate memory capacity, but it must be emphasised that this notion is not fundamental to psycho-logic.

The size of memory span is greatly affected by redundancy of the stimulus material. That is to say, if an item in a list to be recalled is deducible from other items, a subject's memory span will be increased by that redundant item. However, we will be concerned with items that are by definition non-redundant : these G. A. Miller (1956) has named chunks. Each digit in a digit span test forms a chunk, unless the subject is capable of mentally combining the digits into groups. Thus, from the norms of digit span tests (e.g., Wechsler 1949, 1955) we can say that the immediate memory can simultaneously retain two chunks at about 2½ years of age, four by 4½, and seven by 13½, falling to six at about 55 years of age. Thus, immediate memory capacity increases and declines in the same way as intelligence, correlating between ·42 and ·62 with full Wechsler Intelligence Scale scores. The reason why the correlation is no higher may be that, although a given stage cannot arise until the capacity to retain enough chunks has developed, it is not consistently manifested until much later, when the child has learnt to combine and separate the chunks mentally.

Let us call that part of his environment which a subject is considering at any moment his universe. This universe can be characterised by the algebra of classes. In this logical system, any collection of entities, physical or mental, is called a class. A class can consist of a single entity. A class with no members (e.g., the class of unicorns) is said to be empty. There can also be classes formed of classes. If A and B are classes, their logical product is the class of entities in both A and B ; and their logical sum is the class of entities in A or B or both. Classes which are parts of a larger class are called its subclasses. The universe class comprises all the classes relevant to a given proposition. The complement A′ of a class A consists of all entities in the universe class that are not in A.

We may now define a subject's concept of a class as the specification by which he decides whether or not an entity is in that class.* To put it another way, a subject's concept of a class consists in his knowledge of its distinguishing attributes. Gestalt psychology shows that we can form a single unitary concept of an entity or class of entities with any number of attributes. However, the subject, by selectively attending to any combination of attributes, forms as many concepts as there are combinations to consider. For example, if he sees a big white swan, he may apprehend the object either under the concepts 'swan,' 'big (thing),' 'white (thing),' 'big white (thing),' 'big swan,' 'white swan' or 'big white swan.' Furthermore, the subject might restructure the concepts 'swan' and 'black' to form a concept of their product, which would be manifested by some such remark as: "Are there black swans?" He might also restructure the concepts 'swan' and 'raven' to form a concept of their sum, for which he would probably know the name 'bird.' By selective attention to aspects of the concept of a sum or of a product, the subject could again form the component concepts 'swan,' 'black,' 'big,' etc.

---

* Again for simplicity of exposition, it is assumed that a subject categorises a given entity with certainty as either being or not being in a given class. If, however, he categorises an entity as possibly being in a given class, the argument of this paper still stands, but the logical apparatus required is the modal calculus noted by Lewis and Langford (1932, p. 501) as being equivalent to that discussed here.

Now, making a deduction reduces to solving a problem of the type: given that certain classes are or are not empty, what other classes are or are not empty? At the worst, all the classes defined by the attributes of the given classes must be considered. The maximum number of classes that might need to be considered is easily calculated. Suppose the given classes are distinguished by a number, N, of different attributes (or sets of attributes that occur together), then each relevant, distinct class is defined by one of the 2^N possible combinations of the presence or absence of those attributes. For example, the classes given by the premisses of a syllogism vary in three attributes, so finding a conclusion might involve considering up to $2^3=8$ classes. In practice, many problems are solvable without considering all the classes that might be relevant.

As Bernstein (1935) has shown, the importance of N is that for each value it is given, it specifies a unique logic, called a finite Boolean algebra. We will, therefore, label each level of psycho-logic by the appropriate value of N—which must be a positive integer, unless it is 0, as will be explained below.

Because the 2^N classes defined by N attributes are non-redundant, the concept of each class corresponds to a chunk. Therefore, the level at which a subject could reason can be inferred from his digit span, always presuming that this is equivalent to his memory span for concepts: the subject will be at level 1 if the span is 2; at level 2 if it is 3 or 4; and at level 3 if it is from 5 to 8. As already mentioned, finite Boolean algebras would still be valuable as descriptions of the logical status of reasoning at the various psychological stages, even if stages and memory spans did not correspond. But if there were no correspondence, the hypothesis that a certain memory span is a pre-requisite for reasoning at a certain level would have to be discarded. However, a search of the literature, including most of that cited by Munn (1954, p. 412 ff.), has failed to reveal evidence of children making inferences while not having the memory spans required by the hypothesis. Indeed, remembering that memory capacity is only a necessary condition for reasoning at a certain level, and that success cannot be assured without the maximum capacity required at that level, it will be shown that there is a satisfactory agreement between the ages at which the various memory spans are attained and those at which the corresponding levels begin.

The following sketch of levels and stages uses those of Piaget's hypothetical constructs that are clearly supported by empirical data; terms for these constructs are italicised.

LEVEL 0. This demands the capacity to process $2^0=1$ concept at a time. It corresponds to the *stage of sensorimotor intelligence*. This is defined by Piaget (1937) as the period up to two years of age, during which a baby learns to form *object concepts*, that is, to treat a given object as the same object however, and whenever, it is perceived.

We may suppose that some conditioning process leads the baby to categorise identically percepts similar in important respects. It is not known exactly how object concepts are attained. Piaget (1957, pp. 86-9) has suggested that a baby twiddling some object around finds that there is an invariant in the group of spatial displacements he has performed upon the object, and thus learns that the object is permanent : yet it must be pointed out that any finite number of twiddles may not form a group, and how invariants are extracted from perception remains unexplained.

For present purposes, it suffices to remark that the baby can make no inferences. This is due to the fact that an inference, by definition, involves relating at least two concepts, and at level 0 only one concept can be differentiated at a time. It may now seem paradoxical to suggest that the baby's object concept has $N=0$ distinguishing attributes. The paradox vanishes, however, when it is noted that if he has only one concept in mind there is no other concept to distinguish it from, therefore, distinguishing attributes are meaningless in this context.

LEVEL 1. This level demands the capacity to retain $2^1=2$ concepts simultaneously. It corresponds to the *stage of pre-operational intelligence*. This lasts from two to seven years of age, and is defined by Piaget and Inhelder (1941) as the time during which the infant comes to form *class concepts*, that is, to classify objects.

At first the infant may fail to note the distinguishing attributes specified in an adult's concept of a given class, so he may make classifications which adults think mistaken. To overcome this he will, presumably, try to learn the distinguishing attributes of all manner of classes ; this he seems to be doing when he asks a series of questions of the form : " What's that ? A cow ? Why ? What's it for ? "

A more serious source of confusion is the infant's inability to distinguish between more than an object and its environment. He cannot at the same time note the relationship between the object and its environment, because that would necessitate the simultaneous retention of a third concept. But knowing the relative position of an object is essential to deciding whether percepts of similar objects encountered at different times should be classified as a number of objects of the same class, or as the same object. Piaget (1937, p. 374) has observed that infants are often in doubt on this score.

On the other hand, the infant's capacity to retain two concepts at once enables him to draw comparisons, which in turn permits him to make a rudimentary form of inference. If he is confronted with an object he can form and retain a concept specifying those of its attributes to which he has attended. If he now encounters some new object, he can compare it with the retained object. But as he can simultaneously retain two concepts at most, he must classify his concepts of the two objects as being either identically the same, or quite different : he cannot have a third concept of their similarities and differences. Therefore, if the new object has at least one of the attributes of the retained concept, he may expect to find that it has all the other attributes as well. Stern (1924) calls this kind of reasoning from one concrete instance to another 'transduction': it is exemplified by his daughter who, at the age of 3, knew a rope as something to support the family wash, so when she saw a rope strung along a balcony, she asked: "What's that hanging up there? Something to pin clothes on?"

Thinking at this stage is also characterised by the phenomenon of *irreversibility*. It is exemplified by an infant who is shown some brown beads and some white beads. He will correctly tell a questioner that the brown beads are wooden and that the white beads are wooden. But, if he is shown a box containing a larger number of brown than white beads, and he is asked whether it holds more brown beads or more wooden beads, he will insist that there are more brown beads. This would be predicted of an infant who cannot think of three concepts simultaneously: for, if he has already classified the white wooden beads as 'wooden,' and the brown wooden beads as 'brown,' or even 'brown wooden,' he has not the capacity to restructure his concepts to form a third concept 'wooden of any colour.'

LEVEL 2. This level demands the capacity to process up to $2^2=4$ concepts simultaneously. It corresponds to the *stage of concrete operational intelligence*. This is defined by Piaget and Inhelder (1959) as the time, between 7 and 11 years of age, when the child learns how to use *concrete operations*, which are processes of restructuring concepts, that can be illustrated by grouping tangible objects.

Suppose that a child has formed concepts of two classes, A and A′. If he considers the attributes of A and A′ simultaneously, he thereby forms the concept of the sum of A and A′. This new third concept may be designated B. Now the child's immediate memory can accommodate the concept of a further class, B′, which can in turn be coupled with the concept of B. This process, called the concrete operation of *addition*, may be continued indefinitely. It is manifested when the child couples classes of objects to form a wider class. Addition is also manifested when the child puts objects into a series ordered by some specific relation. This is now possible because the child has the capacity, not only to retain concepts of each pair of objects involved,but also the further concept of the relations between them.

The notion that the child forms wider class concepts by combining concepts of subclasses two at a time is similar to that advanced by Piaget (1942; 1947, p. 43; 1949), who states that a child forms the concept of a class (e.g., animals) by combining the concept of a subclass (e.g., insects) with the concept of the complement of that subclass (non-insects). But the complement seems to be defined *with respect to the new class* (thus making 'non-insects' mean '*animals* that are not insects'). This would imply that the child must already have the concept of the new class in order to form that concept—a circularity which obviously cannot be intended.

Next suppose that the child has in his immediate memory the concepts of two classes C and D. Then the child can classify further concepts using, at the same time, the two criteria, 'having the attributes of C' and 'having the attributes of D.' This is manifested in the concrete operation of *multiplication*, that is, in sorting objects into four product classes according to whether they have the attributes of C and D; of C but not of D; of D but not of C; or of neither C nor D. Multiplication is also manifested when a child finds correspondences between two given independent series.

The delay between the attainment of a four-chunk memory span, between the ages of 4 and 5, and the start of the concrete operational stage, at about 7 or 8 years, is more apparent than real. In fact, Piaget and Inhelder (1959) give statistical tables which may be summed up by saying that no 4-year-olds can perform concrete operations, but that some can be carried out systematically by an increasing number of subjects from the age of 5. Evidence of the early performance of concrete operations is also given by Welch (1940) who, incidentally, shows the relation between the abstractness of concepts and their number. Welch distinguishes stages in the development of abstract thought by the hierarchic order of concepts a child can form. At 26 months, on average, his subjects attained first-order concepts : for example, potatoes and apples were both classed as foods. The second stage was reached at 4½ years, when his subjects attained second-order concepts, potatoes being classed as vegetables, apples as fruit, and both fruit and vegetables as foods.

Concrete operations are said to be *reversible*, because the child can behave as if he had mentally cancelled an operation, either by doing the opposite (a type of reverse operation called *inversion*), or by performing an operation that compensates for the first one (*reciprocity*). Thus, for the first time he can argue that if lemonade is poured from a tall, slim glass into a short, fat one, there will still be the same amount to drink, because the lemonade can be poured back again to look as it did at first (inversion), and because decreasing the lemonade's depth has been compensated for by giving it more breadth (reciprocity).

At this stage, however, the child cannot understand how inversion and reciprocity can be combined and so effect *correlation*. For example, the child may be faced with an apparatus consisting of two communicating vessels containing a liquid ; if a piston is pressing on the liquid in one vessel, the level of the liquid in the other vessel will vary according to the weight of the piston and the density of the liquid. A child, at this stage, cannot educe the correlate needed to answer the problem : " The operation of increasing the weight of the piston is to decreasing the weight of the piston (inversion) as increasing the density of the liquid (reciprocity) is to . . . what ? " (Inhelder and Piaget, 1955, p. 148 ff.). The hypothesis of an immediate memory limited to holding four concepts at a time accounts for the child's inability to educe such a correlate, for to understand the problem he would have to retain not only the concepts of an operation, its inversion and its reciprocity, but also concepts of the three relationships involved.

The hypothesis also explains why it is not until the next stage that a child can remember to keep all factors constant except one during an investigation (Inhelder and Piaget, 1955, p. 46 ff.).

LEVEL 3. This level demands the ability to process up to $2^3=8$ concepts simultaneously. This corresponds to the *stage of formal operational intelligence*, which is reached by 11 to 15 years, when the adolescent thinks in conformity with a fully practical logic.

If the adolescents can now retain up to eight concepts, he can, in principle, make any inference involving classes distinguished by up to three attributes (or sets of attributes), such as a syllogism, so the limitations of the previous level vanish for nearly all practical purposes.

Indeed, it will be an unusally complex problem that demands the retention of as many as eight concepts. The conclusion to the simplest form of syllogism (known as Barbara) can be found even with level 2 logic by a process akin to logical addition : the first premiss says that all of class A has the attributes of class B, so that the concept of their sum, A and B, can be formed and used to substitute for A or for B ; the second premiss says that all of class B has the attributes of class C, so that class A has the attributes of class C, which is the required conclusion. However, if this kind of reasoning is applied to other premisses, it may produce invalid conclusions. In fact, most fallacies in reasoning can be accounted for on the hypothesis that the problem being considered calls for the construction of more concepts than can be kept distinct in the immediate memory.

A level 4 logic, which would enable subjects to solve problems involving concepts varying in four different ways, is, of course, possible. However, it is a rare person who is intelligent enough to solve such problems mentally, because they require the simultaneous retention of up to $2^4 = 16$ concepts.

Acknowledgements.—Thanks are due to Professor E. A. Peel and Dr. A. R. Jonckheere for sugggesting improvements to an earlier draft of this paper.

References.

(Page numbers in the text refer to English translations where possible)

Bernstein, B. A. (1935). On finite Boolean algebras. *Amer. J. Math.*, 57, 733-742.

Beth, E. W. (1950). A propos d'un " Traité de logique." *Methodos*, 2, 258-264.

Grize, J. B. (1960). In *Etudes d'épistémologie génétique, XI—Problèmes de la construction du nombre.* Ed. J. Piaget. Paris : Presses universitaires.

Inhelder, B., and Piaget, J. (1955). *De la logique de l'enfant à la logique de l'adolescent.* Paris : Presses universitaires. Trs. A. Parson and S. Milgram (1958). *The Growth of Logical Thinking.* London : Routledge and Kegan Paul.

Langer, S. K. (1953). *An Introduction to Symbolic Logic,* 2nd ed. New York : Dover.

Lewis, C. I., and Langford, C. H. (1932). *Symbolic Logic,* 2nd ed. New York : Dover.

Miller, G. A. (1956). The magical number seven, plus or minus two : some limits on our capacity for processing information. *Psychol. Rev.*, 63, 81-97.

Munn, N. L. (1954). In *Manual of Child Psychology*. Ed. L. Carmichael. New York : Wiley.

Piaget, J. (1937). *La construction du réel chez l'enfant.* Neuchatel : Delachaux et Niestle. Trs. M. Cook (1955). *The Child's Construction of Reality.* London : Routledge and Kegan Paul.

Piaget, J. (1942). *Classes, relations et nombres.* Paris : Vrin.

Piaget, J. (1947). *La Psychologie de l'intelligence.* Paris : Colin. Trs. M. Pearcy and D. E. Berlyne (1950). *The Psychology of Intelligence.* London : Routledge and Kegan Paul.

Piaget, J. (1949). *Traité de logique : Essai de logistique opératoire.* Paris : Colin.

Piaget, J. (1952). *Essai sur les transformations des opérations logiques.* Paris : Presses universitaires.

Piaget, J. (1953). *Logic and Psychology.* Manchester : University Press.

Piaget, J. (1957). In *Etudes d'épistémologie génétique, II—Logique et équilibre.* Ed. J. Piaget. Paris : Presses universitaires.

Piaget, J., and Inhelder, B. (1941). *Le développement des quantités chez l'enfant.* Neuchatel : Delachaux et Niestle.

Piaget, J., and Inhelder, B. (1959). *La genèse des structures logiques élémenataires.* Neuchatel : Delacahux et Niestle.

Stern, W. (1924). *Psychology of Early Childhood.* New York : Holt.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children* (Manual). New York : The Psychological Corporation.

Wechsler, D. (1955). *Wechsler Adult Intelligence Scale* (Manual). New York : The Psychological Corporation.

Welch, L. A. (1940). A preliminary investigation of some aspects of the hierarchical development of concepts. *J. Gen. Psychol.*, 22, 359-78.

(*Manuscript received, 9th September,* 1962.)

PREDICTING STUDENT PERFORMANCE IN A UNIVERSITY DEPARTMENT OF EDUCATION

BY F. W. WARBURTON, H. J. BUTCHER AND G. M. FORREST
(*Department of Education, University of Manchester*)

SUMMARY. One hundred teachers in training in the Department of Education, University of Manchester, were tested with measures of abilities, personality, interests, values and general culture, and provided a considerable range of biographical information. The predictive power of these various measures was examined in terms of three main criteria—final teaching mark, final theory mark, and award of the Graduate Certificate in Education. Degree class was the best single predictor of theory mark and of final award. A personality questionnaire (Cattell's 16 P.F.) was the best single predictor of teaching mark. The successful student tended, apart from academic qualifications, to have a high level of general culture, high conscientiousness (16 P.F., factor G), to have tender-minded attitudes to education, and to participate fairly widely in social activities.

I.—INTRODUCTION.

THE literature dealing with the prediction of success in teaching and success in teachers' training courses is very large and very inconclusive. Some idea of its extent may be obtained from the annotated biography of Domas and Tiedeman (1950) which contains over 1,000 references to papers on 'teacher competence' up to 1950. Its inconclusiveness is also well known. As Evans (1959) says: "The most obvious deduction from the available evidence is that there is no one pattern of successful teacher. Teaching is a complex process calling for many different abilities, no one of which is by itself sufficient to ensure success. At the same time, each of these abilities makes a small but necessary contribution to success" (p. 33).

Among the variables which have been previously studied as possibly related to teaching success are: biographical data such as age, marital status (e.g., Ryans, 1951), intelligence and abilities (in a large number of studies), personality factors measured by questionnaire or by objective tests (Lamke, 1951; Schwartz, 1950), interests (Strong, 1943), attitudes to teaching (Cook, Leeds and Callis, 1952) and many others. In addition to the study of theoretical predictors of this type, much research has been devoted to the predictive power of interviews and kindred methods of selection (e.g., Burroughs, 1958).

The results have, in general, been rather disappointing. Several causes have clearly contributed to the prevailing uncertainty in this field. Firstly, as stated by Evans in the quotation above, the criterion is a complex one. Secondly, in some studies the criterion has been teaching mark at student level, in others rated success as a teacher after varying periods of service. Little is known about the relation between these two criteria, as is pointed out by Collins (1959), since the few correlations available range from + ·84 (Tudhope 1942) to + ·16 (Martin, 1944). Thirdly, the numerous variables which have been tested as predictors have generally been tried out on quite different samples, varying in age, nationality, sex, degree of selection and other respects, with the result that it is hard to compare the results obtained and the relative values of the different types of predictor.

These last two factors are our justification for the present paper, for it describes the administration of tests and the collection of data of all the types described above with one sample of students in a typical university department of education. Although the present research uses the teaching practice marks of graduate students as its principal criterion, it is also planned to follow up the sample after five years' teaching experience, and thus, firstly, to provide further evidence about the relation between practice marks and later success, but also, more importantly, perhaps, to determine which of the predictors of student teaching success hold up, and which deteriorate, when considered against the later and more fundamental criterion of success in teaching as a career.

II.—DETAILS OF THE RESEARCH.

(*a*) *Objectives.*

The programme of research thus has three main aims: (i) to obtain as complete a picture as possible of a sample of graduates entering the teaching profession in terms of biographical and academic background, abilities, interests, values and personality factors; (ii) to determine the relative efficiency of these variables in predicting success in the course of professional training; (iii) to follow up the subjects after five years' teaching experience and re-determine the relative predictive values in terms of success in teaching as a career. The present paper is mainly concerned with the second of these aims.

(*b*) *The subjects.*

Students taking the Graduate Certificate in Education course in the Department of Education, University of Manchester, in the session 1957-58, numbered 133. Volunteers were called for to undergo an extensive programme of psychological testing, and 118 attended the first period of testing, but inevitably by the time data on approximately 100 variables had been assembled, a number of gaps had appeared. The number of students on whom information on every variable was available had shrunk to rather less than 100. However, to facilitate analysis by the electronic computer, and to avoid the complication of slightly different numbers in the sample in different variables, the gaps were filled by assigning to a student who had missed a test, or about whom information was missing, the mean score of the whole group on that particular variable. This expedient was necessary in 2·3 per cent. of the entries in a 100×100 score matrix.

(*c*) *The variables.*

The variables included in the research comprise fifty-seven test scores, twenty-five biographical particulars and eighteen criteria of success on the training course, as follows:

TABLE 1

(1) TEST SCORES.

1—Non-verbal intelligence.	Culture Free Scale 3A (Cattell).
2—Verbal Intelligence.	National Institute of Industrial Psychology. Group Test 33.
3—Spatial ability.	National Institute of Industrial Psychology. Group Test 80A.
4—Memory.	Primary Mental Abilities. Thurstone.
5—Verbal Fluency.	Primary Mental Abilities. Thurstone.
6—Mechanical Comprehension.	Mechanical Comprehension. Bennett.

7—Vocabulary.	Primary Mental Abilities. Thurstone (70 items), together with 30 difficult words from various other sources.
8—General Culture.	100 item test, constructed by H. J. Butcher and F. W. Warburton, Department of Education, University of Manchester. The subject is required to give the name of the authors of a list of books, covering the humanities, sciences and social sciences and works of art and music.
9—Cyclothymia *vs.* Schizothymia.	16 P.F. Factor A (Cattell)
10—Intelligence.	,, ,, Factor B ,,
11—Ego Strength.	,, ,, Factor C ,,
12—Dominance.	,, ,, Factor E ,,
13—Surgency.	,, ,, Factor F ,,
14—Conscientiousness.	,, ,, Factor G ,,
15—Extraversion.	,, ,, Factor H ,,
16—Sensitivity	,, ,, Factor I ,,
17—Paranoid Tendency.	,, ,, Factor L ,,
18—Unconventionality.	,, ,, Factor M ,,
19—Sophistication.	,, ,, Factor N ,,
20—Guilt Proneness (Anxiety).	,, ,, Factor O ,,
21—Radicalism	,, ,, Factor Q_1 ,,
22—Self-sufficiency.	,, ,, Factor Q_2 ,,
23—Self-control.	,, ,, Factor Q_3 ,,
24—Id Pressure (Tenseness).	,, ,, Factor Q_4 ,,
25—Outdoor Interests.	Kuder Preference Record–Vocational Form CH.
26—Mechanical Interests.	,, ,, ,, ,,
27—Computational Interests.	,, ,, ,, ,,
28—Scientific Interests.	,, ,, ,, ,,
29—Managerial Interests.	,, ,, ,, ,,
30—Artistic Interests.	,, ,, ,, ,,
31—Literary Interests.	,, ,, ,, ,,
32—Musical Interests.	,, ,, ,, ,,
33—Social Interests.	,, ,, ,, ,,
34—Clerical Interests.	,, ,, ,, ,,
35—Theoretical Values.	Study of Values (Allport, Vernon and Lindzey).
36—Economic Values.	,, ,, ,, ,,
37—Aesthetic Values.	,, ,, ,, ,,
38—Social Values.	,, ,, ,, ,,
39—Political Values.	,, ,, ,, ,,
40—Religious Values.	,, ,, ,, ,,
41—Radical attitudes.	Inventory of Social Attitudes (Eysenck).
42—Tenderminded attitudes.	,, ,, ,, ,,
43—Radical attitudes to education.	Survey of Opinions about Education (Oliver), Department of Education, University of Manchester.
44—Tenderminded attitudes to education.	,, ,, ,, ,,
45—Naturalist attitudes to education.	,, ,, ,, ,,
48—Accuracy in non-verbal intelligence test.	Proportion answers correct in Test 1.
49—Accuracy in verbal intelligence test.	Proportion answers correct in Test 2.
50—Accuracy in spatial ability test.	Proportion answers correct in Test 3.
51—Accuracy in mechanical ability test.	Proportion answers correct in Test 6.
52—Accuracy in vocabulary test.	Proportion answers correct in Test 7.
53—Accuracy in general culture test.	Proportion answers correct in Test 8.
54—Emphasis in personality test.	No. ' extreme ' answers in 16 P.F. Test.
55—Emphasis in values test.	No. extreme answers in study of Values test.
56—Emphasis in social attitudes test.	No. extreme answers in Inventory of Social Attitudes.
57—Emphasis in radical attitudes to education.	No. extreme answers (radical) Survey of Opinions about Education.

58—Emphasis in tender-minded attitudes to education.	No. of extreme answers (tender-minded) Survey of Opinions about Education.
59—Emphasis in naturalist attitudes to education.	No. extreme answers (naturalistic). Survey of Opinions about Education.

(2) BIOGRAPHICAL PARTICULARS.

60—No. of G.C.E. ' O ' or ' A ' subjects.	No. of subjects passed either at ' O ' or ' A ' level in General Certificate of Education, i.e., subjects in which at least ' O ' level is reached.
61—No. of G.C.E. 'A' subjects.	No. of subjects passed at 'A' level in General Certificate of Education.
62—Degree class.	First class honours=5. Second class honours (upper division)=4. Second class honours (lower division)=3. Third class honours=2. Pass degree=1. No degree=0.
63—Grant.	State scholar=2. Otherwise grant aided=1. No grant=0.
64—Previous teaching experience.	Questionnaire. Any form of teaching, e.g., supply, Sunday School, armed forces=1. No teaching=0.
65—Degree course tutor's reference.	A–E scale.
66—Interview grade.	Mean grade of 2–3 individual interviewers working separately. A–E scale.
67—Social status of school.	Independent School=2. Direct Grant=1. Grammar School=0.
68—Age of university.	Classified as follows : Oxford=2. Cambridge=2. London=1. Manchester=1. Liverpool=1. Leeds=1. Sheffield=1. Nottingham=1. Durham=1. Southampton=0. Exeter=0. North Staffordshire=0. Aberystwyth=0.
69—Arts Faculty.	Arts=2. Social Science=1. Science=0.
70—Age.	
71—Male Sex.	
72—No. of University offices.	Questionnaire. No. given.
73—No. of University societies.	Questionnaire. No. given.
74—No. of interests.	Questionnaire. No. given.
75—No. of games.	Questionnaire. No. given.
76—University residence.	University Hall of Residence=2. Lodgings=1. Home=0.
77—Father's education.	University=5. Training College=4. Public School=3. Grammar School, High School=2. Technical School, Commercial School, Trade School=1. Elementary School=0.
78—Mother's education.	As ' Father's education.'

79—Parent's occupational level.	Grades I–V. Classification of Occupations (H.M.S.O.).
80—Large family size.	No. of children.
81—Late position in family.	1st, 2nd, etc.
46—Religious affiliation.	Questionnaire. Fundamentalist Sects=5. Nonconformists=4. Roman Catholics=3. Church of England=1. No religion=0.
47—Right political affiliation.	Questionnaire. Conservative=4. Liberal=3. Labour=1. Communist=0.
99—Entry grade.	A–E scale. A mark given before entrance to the Department of Education. It is based on the applicant's complete record, including examination, tutor's reference, interview grades and life record.

(3) Criteria of Success.

91—Final Teaching Mark.	A–E scale. Graduate Certificate in Education examination.
97—Final Theory Mark.	,, ,, ,, ,,
98—Final Certificate Award.	,, ,, ,, ,,
82—Tutor's Teaching grade (Michaelmas Term)'	A–E scale by tutor supervising teaching practice.
83—Tutor's teaching grade (Lent Term).	A–E scale by tutor supervising teaching practice.
84—Schools' teaching grade (Michaelmas term).	A–E scale by headmaster of school in which student's school practice was carried out.
85—School's teaching grade (Lent term).	,, ,, ,, ,,
86—Tutor's teaching check list (Michaelmas term).	Check list of 43 items concerned with student's disciplinary control, lesson construction, attitude to pupils, etc.
87—Tutor's teaching check list (Lent term).	,, ,, ,, ,,
88—Extra Tutor's teaching check list.	,, ,, ,, ,,
89—School's teaching check list (January).	,, ,, ,, ,,
90—School's teaching check list (April).	,, ,, ,, ,,
92—Psychology of education examination.	A–E scale. Graduate Certificate in Education examination.
93—Philosophy of education examination.	,, ,, ,, ,,
94—History of education examination.	,, ,, ,, ,,
95—Health education examination.	,, ,, ,, ,,
96—Essay.	,, ,, ,, ,,
100—Curriculum and Method examination.	,, ,, ,, ,,

Scores on the 100 variables were transformed into normalized T scores. The resulting 4,950 product moment correlations were calculated, on the Manchester University electronic computer. Several factor analyses were also carried out, which it is hoped to report in a later paper.

(*d*) *Particulars of the main criteria.*

The prediction of teaching success to be reported in the next sections will be in terms of three criteria, (*a*) final teaching mark, (*b*) final theory mark, (*c*) award or non-award of the certificate. All three refer to the Graduate Certificate in Education of the University of Manchester. In general, the process of arriving at these marks and at the award is believed to be fairly typical of the procedure in other university departments of education.

Final teaching mark is based primarily on a tutor's observations and assessment of a student during about twelve weeks of school practice, although reports from the schools and from an external examiner will, in many cases, be bound to affect it. Final theory mark is based on the average of five examination papers and one long essay. Award of the Graduate Certificate represents the decision of a group of examiners on receiving reports of final teaching mark and final theory mark, together with the testimony of examiners who know the student.

III.—THE PREDICTION OF SUCCESS AT TEACHING.

(*a*) *Prediction from individual measures.*

All significant correlations obtained between any of the tests or biographical variables and the three main criteria are shown in Table 2 below :

TABLE 2

SIGNIFICANT CORRELATIONS BETWEEN THE PREDICTORS AND THE MAIN CRITERIA OF SUCCESS.

Predictors	Criteria		
	Final Teaching Mark	Final Theory Mark	Certificate award
TESTS :			
7—Vocabulary	—	·236	—
8—General Culture	—	·399*	·308*
11—Ego strength (16 P.F.)	—	·194	—
14—Conscientiousness (16 P.F.)	·236	·236	·198
16—Sensitivity (16 P.F.)	·220	—	·215
18—Unconventionality (16 P.F.)	—	·199	—
23—Self-Control (16 P.F.)	·223	—	—
36—Economic Values (Vernon)	—	— ·213	— ·257*
44—Tendermindedness in education	—	·199	·210
49—Accuracy (verbal intelligence)	—	·203	—
52—Accuracy (vocabulary)	—	·258*	·231
54—Extreme responses to personality questionnaire	—	— ·275*	— ·204
57—Emphasis on scale of radicalism in education	—	·262*	·246
BIOGRAPHICAL PARTICULARS :			
60—G.C.E. ' O ' or 'A'	—	·244	·305*
62—Degree class	·238	·569*	·514*
63—Grant	—	·257*	·231
65—Degree Course, Tutor's reference	·249	·366*	·415*
66—Interview grade	·265*	·289*	·254*
72—No. University offices held	—	·242	—
74—No. of interests	—	·261*	·263*
99—Entry grade	·276*	·358*	·322*

All the correlations in the above table are significant at the ·05 level, except those marked with an asterisk, which are at the ·01 level.

(i) *Prediction from Psychological tests.*

Table 2 shows that of the 171 correlations between test scores and the three main criteria, only twenty-two (13 per cent.) are significant. This may seem a small number, but it must be remembered that we would not necessarily expect a large proportion of variables to be related to success in teaching, or in any other particular profession. As regards the type of test, significant correlations were obtained for 17 per cent. of the cognitive test scores, 18 per cent. of the personality measures, 0 per cent. of the interests, 10 per cent. of the values and 15 per cent. of the attitudes assessments (22 per cent. for attitudes to education). There is an important reason, however, why the size of the correlation coefficients should not be taken quite at its face value. This is the familiar problem of correction for range. Although it is impossible to state the extent to which our sample of graduate students forms a selected group compared with the total group of successful and unsuccessful applicants for entry to the course, since the degree of selection and consequent narrowing of range will be different for every variable, there is little doubt that if the tests and predictors had been administered to all candidates, the majority of the correlation coefficients would have been markedly higher, and many of those that are at present only suggestive would become statistically significant. A striking example of the effects of correction for selection in a rather similar situation is provided by Vernon's (1950) study of Civil Service Selection procedure, in which all the separate, uncorrected validities are extremely low, but, when corrected for selection and combined, yield a considerable degree of prediction. On the whole, the results indicate the relative success of two single long tests, Cattell's 16 P.F. personality test, and Oliver's "Survey of Opinions about Education," and the relative failure of the cognitive tests and the measures of interests. For success on the theoretical side of the course, culture and vocabulary are important, suggesting that Arts students, with their literary and verbal background, have an advantage in this respect. Ego-strength, conscientiousness and lack of extreme personality characteristics are also important, perhaps in the preparation of examinations. There is also some indication that tender-mindedness, radical attitudes and originality (unconventionality) are a factor in examination success, i.e., in pleasing the examiners.

The 'General Culture Test' emerges as a better predictor of theory mark than any other single test. This unpublished test (Butcher and Warburton) is *prima facie*, a rather crude measure of extent of reading. It consists of 100 items, the large majority of these being the titles of books, with the respondent being required to name the author. In the remainder of the items, the same principle is applied to well-known paintings and musical compositions.

Turning to the variables that do *not* significantly predict theory mark, it is noticeable that all six of the ability tests (variables 1-6) fall into this class, presumably because of the high general intellectual level, and hence the narrow range, of the sample. It is worth noting that whereas one at least of each type of test correlates with the Final Theory Mark (and hence with the Certificate Award), only personality scores are related significantly to practical teaching. The finding that practical teaching is harder to predict than theoretical success is typical of research in this area. The personality qualities related to Practical Teaching are self-control (Q3), conscientiousness (G) and sensitivity (I). This finding is in agreement with an American research by Erickson (Barr, *et al.*, 1961), who administered the 16 P.F. to a group of sixty-four teachers in Wisconsin High Schools. Erickson found factors Q_3, G and M– to be related to teaching success, with obtained correlations which were all of the order ·2 to ·3. The first two qualities are aspects of the major second order personality factor of stability, while sensitivity is related to tendermindedness, an attitude which is shown throughout this study to be positively correlated with success on this training course. This is not a surprising result in view of contemporary trends in education.

Positive correlations between factors G and Q_3 and teaching success are reported in the 16 P.F. handbook, but the correlation with factor I (Emotional Sensitivity) is a new finding which may, however, be related to cultural differences between the U.S.A. and Britain. Certainly, both subjective observation and empirical research (e.g., Cattell and Warburton, 1961) suggest that this quality of sensitivity or tendermindedness is both more highly valued by and more typical of the European than of the North American culture-pattern. One value of the 16 P.F. test in the present context appears to be that it alone, among the wide range of tests used, predicts teaching practice better than teaching theory.

Besides the significant correlations, it is interesting to note those that, perhaps against expectation, are of negligible size. In particular, verbal and non-verbal intelligence are irrelevant to teaching success in this sample, except in so far as non-verbal intelligence shows a *negative* (but non-significant) correlation of −·125. Similarly, the extensive series of scores provided by the Kuder Preference Record and the Allport-Vernon measures of values fail to yield any significant correlation, although one or two are suggestive of some relationship.

Non-significant but suggestive correlations with teaching mark are shown below :

Age	+·173
No. of G.C.E. passes	+·172
No. of interests (as stated on application form for entry to course)	+·166
Family size	+·165
Accuracy on vocabulary test	+·165
Verbal fluency	+·155
Naturalism in education	−·151
Paranoia or suspiciousness (16 P.F.)	−·142
Social interests (Kuder)	+·138
Economic values (Allport-Vernon)	−·136

In general, these correlations, although small, make quite good psychological sense and require little explanatory comment. The concept of ' naturalism ' in education (Butcher, 1959 ; Oliver and Butcher, 1962) refers approximately to a ' child-centred,' ' progressive ' grouping of attitudes. A teacher scoring high on the scale of naturalism would be expected, for instance, to approve and follow many of the precepts of John Dewey.

(ii) *Prediction from biographical data.*

Twenty-five per cent. of the correlations between biographical particulars and the three main criteria are significant as compared with 13 per cent. for tests. As with the psychological tests, examination success is easier to predict than practical teaching. Academic capacity is important, but there is also a relationship with breadth of interests (number of interests and university offices held). Two points of interest may be mentioned about these results in respect of theory marks. Firstly, the correlation between degree class and theory mark is remarkably high, particularly when one considers the necessarily coarse grading in which degree class has to be expressed. From the point of view of predicting theory mark (and, indeed, award of the certificate), degree class is much more efficient than entry grade, although all available evidence is taken into account in arriving at the latter.

Secondly, this is also true of tutor's reference. Both theory mark and award of the certificate are better predicted by the report on a student's performance as an undergraduate than they are by entry grade, although the latter includes knowledge of the former. Degree class is already known to the interviewer, and entry grade is based on all the information available to the interviewer, on the results of the interview, and on any further information that may have subsequently become available, such as additional references. Practical teaching ability is related to four variables but these essentially comprise academic level only.

Putting the test scores and biographical data together, the good classroom teacher appears to be essentially stable, sensitive and of good academic capacity. Students successful in examination have the same characteristics, together with good verbal capacity and tenderminded and radical attitudes to educational problems.

(*b*) *Multiple Correlations.*

Turning to the multiple correlations (Table 3), we see that the personality variables are the best set of predictors for practical teaching, the theory examination and the final certificate award. These variables are all derived from a single test, Cattell's 16 P.F., which can be administered in 40 minutes. It is, of course, commonplace knowledge that personality is of a very considerable importance, and possibly of paramount importance, in determining academic capacity. The present results suggest that it is no longer necessary to neglect personality test scores in predicting student performance. Interest tests scores are also more efficient predictors than the eight ability tests. The test of general attitudes was understandably less effective than the test of educational attitudes. There is a tendency, but it is not very marked, for the multiple correlations to be higher for the theory examination than for practical teaching.

An interesting finding was that, on the whole, twenty-one biographical particulars afforded better prediction (·871) than forty-two psychological variables (·639). It should, of course, be noted that in practice beta weights on different samples tend to fluctuate wildly and multiple correlations fall considerably when the same weights are applied to later samples.

TABLE 3

MULTIPLE CORRELATIONS.

	Practical Teaching	Theory Examination	Final Certificate Grade
8 ability tests	·270	·466	·335
15 personality variables	·438	·538	·448
10 interest scores	·330	·293	·268
6 values	·167	·265	·298
2 attitudes	·084	·170	·134
3 educational attitudes	·207	·241	·257
42 psychological variables	·653	·696	·639
21 biographical particulars	·501	·741	·871

IV.—IMPLICATIONS FOR SELECTION PROCEDURES.

Although considerations of space prevent a full discussion of the several factor analyses carried out, it will be convenient simply to list the seventeen main factors which emerged in the most satisfactory of the analyses and to discuss the most promising predictive measures under these headings. One factor loaded almost exclusively on criterion variables and was interpreted as teaching ability. The remaining sixteen were interpreted as follows:

(1) Academic success.*
(2) Verbal ability.
(3) Non-verbal ability.
(4) Stability of personality.*
(5) Extraversion.
(6) Radicalism.
(7) Tender-mindedness.*
(8) Emphasis in expressing opinions.
(9) Interest in arts subjects *versus* science subjects.
(10) Aesthetic interests.
(11) Participation in social activities.*
(12) Socio-economic status.
(13) Family size.
(14) Assessment on entry to Department.*
(15) Verbal fluency.
(16) General culture.*

Six of these factors were loaded by variables which appeared to have a promising relation with the main criteria, and these have been marked by asterisks in the above list. They are also shown, with the more promising test and biographical variables, in Table 4 below, which includes the combined predictive effects of the groups of variables. The correlations would probably be appreciably higher if the technique of multiple correlation had been used in place of the simple pooling with unit weights.

TABLE 4

TESTS TO BE USED IN SUGGESTED SELECTION PROCEDURE.

	Theory Marks	Practical Teaching	Certificate Award
(1) ENTRY ASSESSMENT.			
Final entry grade	+358	+276	+322
Tutors' reference (degree course)	+366	+249	+415
Total (unit weights)	+408	+296	+416
(2) ACADEMIC SUCCESS.			
University degree class	+569	+238	+514
Total (unit weights)	+569	+238	+514
(3) GENERAL CULTURE.			
General Culture (Butcher and Warburton)	+399	+005	+308
Vocabulary (Thurstone)	+236	+054	+156
Total (Unit weights)	+265	+128	+299
(4) STABILITY.			
Conscientiousness (16 P.F.G.)	+131	+236	+198
Self-control (16 P.F.—Q3).	−101	+223	+063
Total (Unit weights)	+018	+281	+160
(5) TENDER-MINDEDNESS.			
Sensitivity (16 P.F.—I).	+156	+220	+215
Tendermindedness in Education (Oliver)	+048	+199	+210
Total (Unit weights)	+231	+179	+277
(6) SOCIAL ACTIVITIES.			
No. of expressed interests	+261	+166	+263
No. of University offices held	+242	−029	+108
Total (unit weights)	+313	+086	+231
ALL TESTS (unit weights)	+707	+520	+708

In considering these findings, it is also interesting to note the factors to which success in the teaching course is largely unrelated. These are verbal and non-verbal ability, extraversion, radicalism, emphasis in expressing opinions, arts *versus* science interests, aesthetic interests, socio-economic status, family size and verbal fluency.

It would, therefore, appear best in selecting potential teachers to concentrate on attainment and general culture rather than ability, on stability rather than on extraversion or introversion, on academic rather than aesthetic interests, and on participation in social activities rather than on social or domestic background.

V.—REFERENCES.

BARR, A. S., *et al.* (1961). Wisconsin studies of the measurement and prediction of teacher effectiveness. *J. Exper. Educ.*, 30, 1-156.

BURROUGHS, G. E. R. (1958). A study of the interview in the selection of students for teacher training. *Brit. J. Educ. Psychol.*, 28, 1, 37-46.

BUTCHER, H. J. (1959). The attitudes of teachers and student teachers to education. *Unpublished Ph.D. thesis*, Univ. of Manchester Library.

CATTELL, R. B. (1957). *Personality and Motivation Structure and Measurement.* New York: World Books.

CATTELL, R. B., and WARBURTON, F. W. (1961). A cross-cultural comparison of patterns of extraversion and anxiety. *Brit. J. Psychol.*, 52, 1, 3-15.

COLLINS, M. (1959). A follow-up study of some former graduate student teachers. *Brit. J. Educ. Psychol.*, 29, 3, 187-197.

COOK, W. W., LEEDS, C. H., and CALLIS, R. (1952). Manual *of the Minnesota Teachers' Attitude Inventory.* New York: Psychological Corporation.

DOMAS, S. J., and TIEDEMAN, D. V. (1950). Teacher competence: an annotated bibliography. *J. Exp. Educ.* 19, 2, 101-218.

EVANS, K. M. (1959). Research on teaching ability. *Educ. Research* 1, 22-36.

LAMKE, T. A. (1951). Personality and teaching success. *J. Exper. Educ.*, 14, 75-100.

MARTIN, I. O. (1944). *The Prediction of Success for Students in Teacher Education.* New York: Teachers' College, Columbia University.

OLIVER, R. A. C., and BUTCHER, H. J. (1962). Teachers' attitudes to education. I.—The structure of educational attitudes. *Brit. J. Soc. Clin. Psychol.*, 1, 56-69.

RYANS, D. G. (1951). A study of the extent of association of certain professional and personal data with judged effectiveness of teacher behaviour. *J. Exp. Educ.*, 20, 1, 67-77.

SCHWARTZ, A. N. (1950). A study of the discriminating efficiency of certain tests of the primary source personality traits of teachers. *J. Exp. Educ.*, 19, 1, 63-94.

STRONG, E. K. (1943). *The Vocational Interests of Men and Women.* Stamford: Stamford University Press.

TUDHOPE, W. B. (1942 and 1943). A study of the training college final teaching mark as a criterion of future success in the teaching profession. *Brit. J. Educ. Psychol.*, 12, 167-171 and 13, 16-23.

VERNON, P. E. (1950). The validation of civil service selection board procedures. *Occup. Psychol.*, 24, 75-95.

(*Manuscript received, September,* 1962.)

CRITICAL NOTE

SPECIFIC DYSLEXIA

By D. RUSSELL DAVIS* and ASHER CASHDAN†
(*Medical Psychology Laboratory, University of Cambridge.*)

In areas of psychopathology where aetiologies are but imperfectly understood, and where methods of treatment are still a matter of argument, it is not surprising that controversy should rage over appropriate groupings of symptoms and the terminology most suitable for them. One such case is that of specific developmental dyslexia (formerly word-blindness), which under its new name has recently been receiving fresh attention (see Zangwill, 1960 ; Vernon, 1962 ; Miles, 1962). In our opinion, the controversy over the existence of this condition is often based on an imperfect understanding of the cardinal principles of scientific investigation and classification.

A disquietingly large number of children suffer from serious backwardness in reading. Some catch up quickly, others never learn to read well. Some respond to specialist remedial teaching ; others show a temporary improvement, but relapse as soon as the special teaching is withdrawn (see Collins, 1961) ; others show no positive response at all. Some of the sufferers are girls, but more of them are boys. Some are of normal or superior intelligence, far more of them are of low intelligence. Many of them have other educational disabilities ; many also have emotional and family problems. A few of them have unusual patterns of handedness and eye dominance ; a few also are claimed by some investigators to have special perceptual difficulties, or difficulty in associating symbols ; such perceptual difficulties may be limited to the activity of reading, or may be more general.

The above list is not exhaustive, but implicit in it is the problem that concerns us. This is, to what extent are any, or all, of these factors associated with each other. If they are, we need to know which of the associations are causal ones, and in which direction the causal relationship operates. For instance, does backwardness in reading produce emotional disturbance, or do emotional disturbances lead to a refusal to learn to read ? Or are both true ?

If we studied and treated each case of reading disability separately, we might achieve fair success in curing the disability (or removing the impediment to learning). We would not, however, be able to draw any conclusions about the cause of the disability or the mode of operation of the ' cure ,' unless we were able to plan and carry out the treatment experimentally—and a single case is rarely sufficient for a satisfactory experiment. Again, if we gave identical remedial treatment to two groups of children, each having distinct features—such as right-handed and left-handed children, or children of low and high tested intelligence, or simply boys and girls—suppose then that one group improved significantly more than the other, we might conclude that left-handed children, say, have more persistent difficulties than right-handed ones, or that girls recover more easily than boys. Even in such a situation, alternative hypotheses would be tenable, but these alternatives could reasonably be investigated. However, the fact that one group had done better than another would have justified the classification (provided always that the result was a stable one). In other words, it would be reasonable to classify retarded boys separately from retarded girls if the two groups responded differently to treatment. But, if the response of the two groups was not clearly different from each other, no case would so far have been made out for classifying them as separate groups of retarded readers.

---

* Now at the Department of Mental Health, Bristol University.

† Now at the Department of Education, Manchester University.

In fact, if any category is to be established within the group of backward readers, it must, to serve any useful purpose, be demonstrably different in a significant respect from other (or some particular group of) backward readers. The purpose of such a differentiation would be to make easier the task of preventing, curing, or forecasting the course of the condition. The study of symptoms, and their classification, is only useful in so far as it throws light on psychological mechanisms, or is somehow relevant to these three tasks.

If a new category is proposed, we are suggesting then that it must satisfy one of the following criteria. It must be different in respect of aetiology, prognosis, or treatment indicated. If the new category differs only in respect of symptoms, then the differences are either trivial or point to a need for further study of the relevant aetiology, prognosis or treatment.

Let us now, in this light, examine the proposed category of specific dyslexia, and taking each criterion in turn, ask whether the members of this proposed category can be differentiated from other backward readers.

Little is known of the aetiology of any form of reading backwardness. There is almost certainly some relationship with low intelligence, and with social class, but even this is not well documented. Hereditary factors have been suggested by some investigators, but there is no clear-cut evidence. Hallgren (1950) found hereditary factors in almost all cases; this finding is so divergent from those of other workers as to make us hesitant to accept any of his conclusions. Zangwill (1960) found a significant number of affected relations in the families of dyslexics with handedness problems. But his work is also inconclusive. Such studies raise, incidentally, the question of the value of history-taking to establish familial features in pathological groups. Quite simply, if the mother of a child with some pathology is invited to produce relatives with similar histories, she is highly inclined to do this (and at the same time to alleviate her own guilt feelings). The mother of a normal child has no such inclination. An investigation based on history-taking may be of value, however, if it is designed deliberately to show whether the pattern of incidence of affected relatives conforms to the one predicted from a particular hypothesis. Hallgren's hypothesis, for instance, was that the pathogenic gene is dominant, but his method of proving that the observed pattern of incidence does not depart significantly from the predicted pattern was hardly satisfactory.

So much for aetiology. We can do no better on prognosis. As Vernon (1962) has pointed out, we certainly need to know why, in groups of children given remedial teaching, some improve while others do not, but in fact we do not yet know why this happens. Similarly, too, with methods of treatment. Miles (1962) says that "there is broad agreement as to the general type of remedial teaching that is normally called for," and he advocates, amongst other things, emphasis on phonic rather than on look-and-say methods, and the use of all relevant sense-modalities in the learning process. We would agree with his prescription, but still await evidence that it is indicated for any *specific* group of poor readers.

We cannot, therefore, agree that there is any good evidence for Vernon (1962)'s assertion that "there does exist a number of cases who have great difficulty in learning to read, which difficulty seems not to be attributable to environmental factors or to emotional maladjustment." There are many reasons why psychologists and others might wish to spread such a belief. Some of these are related to the clinical management of cases, and others to the desire to support *a priori* theories about neurology and genetics. But there is not much evidence to support the belief.

There are two legitimate ways in which advances may be made in this field. One is by clinical investigation and by testing theories and hypotheses in the clinical situation, or even by developing new insights there. But this has to be followed by the second method, which is to examine the relevance of separable factors (whether aetiological, symptomatic or prognostic), *one at a time*, in specially selected groups of children who are backward in reading.

To sum up, it is legitimate to adopt for research purposes the hypothesis that there is a special class of backwardness in reading, which we may call specific developmental dyslexia, and which has one or more characteristics, and then to carry out experiments to test the hypothesis itself, or, more likely, a deduction from it. Whether such a hypothesis is a promising one at the present time is a matter of opinion. But it would only be legitimate to assert or imply that the existence of this class had been established if evidence could be provided of a quality which deserved general acceptance. We have indicated in this note the sort of evidence which would be required, and we have given our opinion that it has not yet been provided.

REFERENCES

COLLINS, J. E. (1961). The effects of remedial education. *Educational Monographs, University of Birmingham, Institute of Education, No.* IV.

HALLGREN ,B. (1950). Specific dyslexia. *Acta Psychiat. Neurol. Suppl.*, 65.

MILES, T. R. (1962). Specific dyslexia. *Bull. Brit. Psychol., Soc.*, 48, 51-2.

VERNON, M. D. (1962). Specific dyslexia. *Brit. J. Educ. Psychol.*, 32, 143-150.

ZANGWILL, O. L. (1960). *Cerebral Dominance and its Relation to Psychological Function.* Edinburgh : Oliver and Boyd.

SPECIFIC DYSLEXIA

A REPLY TO DR. RUSSELL DAVIS AND MR. ASHER CASHDAN

By M. D. VERNON

It was made clear in my paper, " Specific Dyslexia," that the classification of certain cases of reading backwardness as ' specific dyslexics ' was presented as a hypothesis based upon the association together of certain symptoms and types of behaviour with reading backwardness with above chance frequency, and in the absence of certain other characteristics such as low intelligence and emotional maladjustment (except in so far as the latter developed after rather than before the onset of reading difficulties). There seems to be reliable evidence to support this hypothesis, notably that obtained by Dr. Ingram (to whom Dr. Davis and Mr. Cashdan do not refer). This hypothesis does no more than suggest tentatively what might be the causal factors, which are obviously very obscure. But it is, I suggest, scientifically interesting, and valuable in so far as it may stimulate further research (which may, of course, disprove it). Whether the hypothesis is useful in " the task of preventing, curing or forecasting the course of the condition " is a medical question on which I am not qualified to adjudicate ; though I would at least suggest that it might be desirable to consider whether backward readers who might be ' specific dyslexics ' should be treated in a somewhat different manner from those whose backwardness appears to be due primarily to environmental factors. Incidentally, I think that the term ' specific dyslexic ' may be preferable to that of ' specific dyslexia ' in so far as the latter suggests the existence of a single well-defined clinical entity, which seems improbable.

I entirely agree that the clinical diagnosis of cases of reading backwardness is often woefully inefficient, and I have said so repeatedly. I have also urged that much more thorough and systematic experimental examination is necessary of the psychological processes characterizing backward readers, as well as of their clinical histories. I agree also that all too little is known about prognosis and treatment. But I cannot agree that " the second legitimate way of advance " must be to examine the relevance of separable factors *one at a time*. This method has been utilized again and again in the past with little useful result. What we need to observe now is the nature and frequency of the association between certain aetiological, symptomatic and prognostic factors in backward readers. When these have been established, it may be possible to offer further hypotheses as to causality.

In support of what I have said here, I should like to quote from a paper in a recently published book, *Reading Disability : Progress and Research Needs in Dyslexia*, edited by John Money, Associate Professor of Medical Psychology and Paediatrics at Johns Hopkins University (Baltimore, The Johns Hopkins Press, 1962) : " It is not at all rare in psychological medicine, nor in other branches of medicine, that a disease should have no unique identifiying sign, that uniqueness being in the pattern of signs that appear in contiguity. Out of context, each sign might be encountered in other diseases, or, in different intensities, in the healthy." Specific dyslexia is no exception in this respect.

A BRIEF COMMENT

ON

THE EFFECT OF SOCIO-ECONOMIC STATUS ON THE SCHOLASTIC ACHIEVEMENT OF SIXTH-GRADE CHILDREN, by ROBERT L. CURRY*

By R. R. DALE

(*University College, Swansea*)

The importance of Curry's admirable article increases the danger that his one faulty conclusion may be repeated by research workers without a careful appraisal of the data from which his finding is drawn. His first conclusion (page 48) is that " Socio-economic status seems to have no effect upon the scholastic achievement† of sixth-grade students when the students have high intellectual ability." Unfortunately his data do not warrant such an assumption. Even within his high intellectual ability group, the same trends are apparent as within his low ability group, the upper socio-economic class gaining superior scores in all cases. The lack of statistical significance for these superior scores may be due to the smallness of the samples which the experimenter has selected. We cannot say that there seems to be ' no effect ' when the effect merely fails to reach statistical significance. All we can say is that the effect may have occurred by chance ; we certainly cannot say in this case that it has probably occurred by chance, and to say that it has certainly occurred by chance—which is the equivalent of saying that in this case socio-economic status has no effect—is statistically unthinkable.

There is also a second fact which deserves consideration. In work which produces a patterned result there is a tendency for us to ignore the importance of the pattern when we are considering questions of statistical significance. In Curry's results there are in effect eighteen differences between socio-economic groups to be considered ; of these sixteen are in the ' expected ' direction, and this in itself has its statistical importance. If we take the extreme socio-economic groups, comparing upper with lower, there are nine differences in mean mark to be considered, and all nine are in the ' expected ' direction. The ' t ' ratios, therefore, have a cumulative effect, and although when we compare the attainment levels of different social classes the individual ' t ' ratios are lower in the high intellectual ability groups, we might well conclude that " the trend is continued but lacks statistical significance ; such significance might be obtained in future experiments by increasing the size of the sample." We need to bear in mind, also, that although we have by consent drawn an artificial line between ' significant ' and ' non-significant,' there is in actuality no such barrier, but a gradual transition.

* This *Journal*, May, 1962, p. 46.

† In Reading, Arithmetic and Language.

RESEARCH NOTE

THE RELATIVE PERSONALITY ADJUSTMENT OF SEVERELY DEAF AND PARTIALLY DEAF CHILDREN

By L. R. BOWYER, A. MARSHALL AND K. WEDDELL.

The policy, consequent on improved electronic aids, of placing partially deaf children in ordinary schools, has raised a question of their need for special help with social and emotional problems. One hypothesis is that a partially deaf person may have more of a personality problem than one who is severely deaf.

A pilot study was carried out in Bristol between May and September, 1962, to try out ways of investigating this hypothesis and to see if there was evidence to warrant more time-consuming research. Three groups were studied, each with ten children, respectively severely deaf, partially deaf, and of normal hearing. The age range was restricted to 8–10 years. The groups were matched by paired (or triple) comparison for age, sex and intelligence. Ranges of I.Q., rather than single I.Q. figures were matched, corresponding to the Stanford-Binet and Wechsler levels of Superior, Bright, Good Average, Average, Low Average, and Dull or E.S.N. For this purpose a judgment was made based on recorded test scores, teachers' opinions, and the results of a Goodenough Draw-a-Man test. The severely deaf children (Grade III deafness) were all from one Special School for the Deaf. The partially deaf children (Grade II deafness) were either members of a deaf unit in an ordinary primary school or were individuals in classes of children with normal hearing.

The Criterion of Adjustment.

The criterion of adjustment devised for this study was one of 'adequacy of fantasy,' assessed from a Lowenfeld 'World' and a Lowenfeld Mosaic pattern, with consideration also of qualitative features in the Draw-a-Man test. This criterion was adopted for two reasons. First, the Lowenfeld techniques do not need words either in the instructions or in response, thus equating the communication problem for all groups.

Secondly, a recent American study of deaf children, (Brill, 1960) which used the criterion of how much each child was liked, showed a trend in favour of the severely deaf children, but it was not statistically significant. The authors think that this may have been because they omitted to separate those with deaf parents from those with hearing parents. In addition, however, the criterion of adjustment may have been too simple or one-sided. The present attempt to gain an idea of the inner experience of the children with the aid of projective techniques is aimed at balancing the picture from the other side. In a larger piece of research it would be important to use both ideas. In our pilot study use was made of the Vineland Social Maturity Scale, and we are grateful to Mr. Geoffrey Herbert for helping with several of these and with Goodenough and Mosaic tests. Attention was also given to teachers' knowledge of the children's homes. Only the 'adequacy of fantasy' results will be reported here.

THE SCORING SYSTEM.

The 'World' and the Mosaic patterns were scored in the following way, with one point for each aspect :

SAND TRAY			MOSAICS		
Recognisable Theme	Adequate Content (more than 4 categories in relationship)	Freedom from adverse Clinical signs	Good Gestalt	Freedom from adverse Clinical signs	Total possible Score
+	+	+	+	+	5

The validity of the World 'Clinical signs' (empty, unpeopled, rigid, over-fenced, aggressive) had been established in previous research (Buhler, 1951 ; Bowyer, 1956) and also for the Mosaics (empty centre, frame design, etc., Lowenfeld, 1952).

RESULTS.

The average scores for the severely deaf, normal, and partially deaf children were 4·1, 3·4 and 2·85. With the Sign test of statistical significance (Siegel, *Non-parametric Statistics*, p.68), using the hearing group as a control or norm, the severely deaf children in this study are above average in adjustment (N 10, x=3, p=·05). However, the partially deaf children are not below average to a statistically significant degree.

The results of comparing the two deaf groups, in matched pairs, equated for sex, age, and intelligence, gives a difference in favour of the severely deaf which is not statistically significant, although it appears noticeable to the naked eye, so to speak. A colleague has suggested that a more powerful statistic, such as analysis of variance, might bring out any significance that may be there.

It is possible to obtain a significant difference between the two deaf groups with the Sign test, if one abandons paired comparison, regarding it as having served its purpose in equating the groups. With the scores arranged randomly, p=·002 can be obtained. By pairing (*a*) for intelligence regardless of age or (*b*) for age regardless of intelligence, then probabilities of ·06 and ·05 are obtained. Perhaps this means that mental age is the most important consideration for adjustment in physical handicap, since combining age and intelligence reduces the probability of difference in adjustment, thus :

Severely Deaf		*Partially Deaf*	
J.B.	5+	V.W.	3½
S.B.	5+	V.G.	2
S.H.	3 −	S.R.	4
J.A.	2½ −	C.R.	4
O.C.	5+	G.B.	2
M.H.	5=	S.S.	5
R.G.	4+	M.T.	1
R.R.	5+	P.E.	2½
J.F.	3 −	J.R.	3½
A.R.	3½+	R.M.	1

N 9, x=3, p=·254

In the above table, the children are matched for age, sex, and intelligence. These various results seem to offer some encouragement to continue research on relative differences between the severely and the partially deaf.

THE NEXT STAGE.

Two studies are planned. One is to look at the playground behaviour in those schools with deaf units, to find out how many social contacts take place between partially deaf and hearing children. The other is to proceed with a larger sample of children in a school which teaches both severely deaf and partially deaf children. For it may be that the partially deaf child is less well adjusted only when he is educated as a member of a minority group in an ordinary school for hearing children, i.e., adjustment may be a function of the Special School with its entire staff, buildings, equipment and grounds devoted to helping the handicapped child. The alternative hypothesis would be that even in this favourable environment the partially deaf child by the nature of his handicap, is a ' marginal man ' who fits in neither with deaf nor hearing children.

SUMMARIES OF RESEARCHES REPORTED IN DEGREE THESES

A Survey and Discussion of the Use of Reports to Parents in Scottish and Other Schools.

By F. J. R. HONIGMANN

(*Summary of thesis presented in part fulfilment of requirements for the degree of Ed.B. in the University of Aberdeen in* 1962)

AIM.

To review the practices in the issue of reports to parents; to compare the practice in Scotland with that in selected parts of England, Wales and the United States; and to consider the ideal form of such a report.

DATA.

Thirty-five education authorities in Scotland, twenty-two in England, nine in Wales and six in U.S.A. (total seventy-two) were asked to submit specimens of reports to parents. Twenty-nine from Scotland replied, providing a reasonably representative sample for Scotland. Replies from Wales and U.S.A. (large cities only) were also representative. In England, the responsibility for issue of reports seems generally to be left to individual schools; requests were sent to a sample of English schools but less than half sent specimen reports.

METHOD.

Reports from secondary schools were selected for examination. Points considered were: frequency of issue; format; nature and quantity of information given (examination marks, standards of comparison, ratings, space provided for remarks, aspects of performance assessed).

RESULTS.

Scottish reports tend to be restricted to an impersonal list of examination marks on formal subjects with very little provision for verbal comments or assessment of personality development, in contrast with English, Welsh and American reports.

CONCLUSIONS.

Developments in educational practice have not been reflected in the form of reports to parents in Scotland. Alternative forms which communicate information more effectively and aid parent-teacher collaboration are suggested.

An Evaluation of Readability Formulae in assessing Age-Placement of Text Books.

By MARGARET TAYLOR

(*Summary of thesis presented in part fulfilment of requirements for the degree of Ed.B. in the University of Aberdeen in* 1962)

AIM.

To study the validity of three American readability formulae, Lorge, Dale-Chall and Devereux, as applied to the age-placement of twelve British school reading books graded by publishers as appropriate for children aged 9 to 10 years.

METHOD.

The criterion used for evaluating the formulae was the combined judgement of forty-five experienced teachers. An order of difficulty among the twelve reading books was calculated from the teachers' rankings, and the order obtained from the application of each of the formulae was compared with this criterion. The teachers also assessed other factors which contribute to the age-placement of the books but which are not measured by the formulae, such as format, appropriateness of content, exercises and illustrations.

RESULTS.

The teachers' ratings showed an average intercorrelation of over +0·7 and, their combined judgment was, therefore, accepted as a reasonably consistent criterion. The Lorge and the Dale-Chell formulae showed moderately high correlation with this criterion (+0·7 and +0·6 on rank-order), while the Devereux formula gave a correlation of only +0·5. All three formulae correlated highly with each other (inter-correlations between +0·8 and +0·9). Age placements by the formulae tended to be too high by comparison with teachers' opinions. There was a correlation of +0·3 between the teachers' order of difficulty and their ranking of the merit of the books as school reading texts. Additional comments from teachers and ratings of other aspects were found to add relevant information beyond that given by the formulae.

CONCLUSION.

The value of two of the readability formulae was established, but only within the limits of what these formulae aim to measure. Readability, however, is shown to be only one of several factors relevant to the appropriate age-placement of these school texts.

A Gaelic Translation of the Wechsler Intelligence Scale for Children.

BY MARY MORRISON

(*Summary of thesis presented in part fulfilment of requirements for the degree of Ed.B.in the University of Aberdeen in* 1961)

AIM.

To test the validity of a new Gaelic translation of WISC, with reference to the study of the effect of bilingualism on intelligence.

SUBJECTS.

All Gaelic speaking children in Lewis, born in April to September 1948 (inclusive), aged 11 : 6 to 12 : 0 at the time of testing. Of 130 such children, 128 were tested.

TESTS.

In addition to WISC (Verbal and Performance Scales), two Moray House Verbal Reasoning Tests, Raven's Progressive Matrices, teachers' estimates of school performance, and indices of the occupational status of the homes and of the extent of Gaelic in the linguistic background of the homes.

RESULTS.

The main findings were :

(i) The Gaelic translation of WISC agreed with the Moray House tests and Matrices in showing an average score of 92 or 93 in the group.

(ii) Gaelic children were at a disadvantage on performance tests, and on strictly timed tests.

(iii) The form of WISC sub-tests recommended for use with Gaelic-speaking children are :

Verbal—Comprehension, Similarities and Vocabulary (all in Gaelic), Arithmetic and Digit Span (both in English).

Performance—Picture Completion, Block Design, Object Assembly, Coding and Mazes (all with Gaelic instructions).

The Effects of Treatment in a Special Class in a Junior School on Retardation and Maladjustment with Special Reference to the Concept of Self

BY DAPHNE C. IVES

(*Abstract of thesis submitted for the degree of M.Ed. in the University of Manchester, December*, 1961)

AIM OF THE INVESTIGATION.

This research was undertaken in order to assess the value, to eleven children of junior school age, of a year spent in a special class. The class was attached to a primary school in a Lancashire cotton town which accommodated 350 children. The children had been placed in the class for one or more of the following reasons : they were borderline cases for placement in a school for E.S.N. children ; or, they were of low-average intelligence and severely retarded in one or more of the basic subjects ; or, because of some emotional disturbance, they had been unable to adjust themselves to a normal classroom situation.

Special emphasis was put upon changes in the self-concept of these subjects.

METHOD.

(1) Measurements and evaluations made at the beginning and at the end of the period included Sander's tests of Social Under-Evaluation (SUV) and Non-Social Tendency (NST),[1] Raven's Projection Test,[2] the Bristol Social-Adjustment Guide, forms 1 and 2,[3] together with tests of persistence and level of aspiration designed by the present writer. In addition, a series of controlled observations were made by the writer, using a check list during free-play periods. Items which were checked were subsequently grouped under four separate headings : hostility, social interaction, withdrawal and constructiveness. All of the data so collected were analysed by the Wilcoxon Matched-Pairs Signed Ranks test[4] to establish the direction and significance of changes over the year.

(2) Additional information was provided by (*a*) questionnaires which were completed at the beginning and end of the year on the basis of separate interviews with the children themselves and with their parents or guardians ; (*b*) questionnaires administered to sixty-six children who attended the school to which the special class was attached concerning their attitudes to the children in this class ; (*c*) questionnaires completed both by the special class children and by a control group taken from the remainder of the school regarding their attitudes towards previous teachers ; and (*d*) four sociometric tests, covering the period.

Although none of the data described in the previous paragraph lent itself readily to statistical analysis, the information was found valuable in forming a qualitative picture of the effects of placement in a special class of this sort.

(3) Educational progress was measured by standardized tests of attainment administered at the beginning and end of the year. In spite of what appeared to be marked progress made by the majority of the subjects, however, the results did not lend themselves to statistical comparison owing to the fact that many of the children had failed to make any score on entry.

RESULTS.

(1) Significant changes were found to have occurred in the areas of adjustment measured by a number of the tests used. Social under-valuations had decreased $(P < \cdot 01)$ as also had non-social tendencies $(P < \cdot 025)$. Although the total number of adverse pointers listed in the Bristol Guides showed no significant change, there was a significant reduction in items classified U/W (unforthcomingness—withdrawal) and K (Knavery or psychopathic hostility to the institutions of society and its representatives as they impinge on the child), $(P < \cdot 025$ in both cases). The subjects displayed a more realistic approach towards future performance, as shown by a decrease in the final discrepancy scores calculated from the test of level of aspiration $(P < \cdot 025)$. Observed incidence of hostility showed a marked decline during the second half of the period $(P < \cdot 01)$ and constructiveness showed a corresponding increase $(P < \cdot 01)$.

No significant changes appeared in ' conformity ' as shown by Raven's Projection Test,[2] in ' persistence,' or in the observed frequency of interaction or incidents classified as ' withdrawn ' during the controlled observations.

(2) The qualitative results of the experiment may be briefly listed as follows :

(*a*) The responses to the questionnaires given to the subjects indicated that they all had a poor self-view in relation to school work at the beginning of the year. Ten of the subjects showed an improvement in this respect by the end of the year.

(*b*) The results obtained from the initial questionnaire given to a parent or guardian of each of eleven subjects, showed that ten of the eleven adults were prepared to co-operate with the special class teacher and were grateful that special help was being provided for their children. At the end of the year, eight of the parents were very satisfied with the social and academic improvements made by their children. One (a guardian) was not interviewed again and the other two had certain reservations about the class, but were not wholly dissatisfied.

(*c*) The responses to a questionnaire given to sixty-six children in other classes within the school indicated that some of them regarded the special class children as being inferior to themselves. Remarks to this effect were made in the playground which could have had an adverse effect on the self-concepts of the subjects.

(*d*) When the subjects were questioned about their attitudes to teachers, the incidence of self-referential statements was higher among the special class children in both the ' Teachers Liked ' and the ' Teachers Disliked ' categories than among the ' normal ' children. This indicates that the self-forming influence of teachers is greater with extremely unsuccessful children.

(*e*) The sociometric inquiries showed that of three children who were socially isolated within this group at the beginning of the year, one had become satisfactorily integrated by the end of the year.

(*f*) Attainment tests indicated that ten of the eleven subjects had made sufficient academic progress by the end of the year to enable them to feel that they were beginning to acquire the basic skills.

CONCLUSIONS.

It should be emphasised that because of the small sample and the limitations of the experimental situation, the following conclusions need to be regarded as tentative.

1.—During the year of the experiment the subjects who had been designated ' withdrawn ' (criterion : ten or more U/W pointers on the Bristol Guides) became significantly less so. It can be tentatively concluded, therefore, that relative success, or a permissive classroom climate, or both, help such recessive children to make a more adequate social-adjustment.

2.—There was a decrease in hostile behaviour and an increase in the number of constructive interactions observed in the second half of the year. This suggests that providing opportunities for free interactions to children who have failed to make a good social-adjustment is beneficial, in so far as it helps them to function more adequately in a group.

3.—After a year in the special class, the subjects tended to make their predictions on the level of aspiration test more in accordance with reality. As the level of aspiration situation is directly related to feelings of success and failure (a high discrepancy score indicating feelings of inadequacy) it is concluded that the decrease in scores is indicative of an improvement in the self-concepts of the subjects which could be attributed to one or more of the following reasons: (*a*) most of the subjects had begun to master the basic skills of reading and arithmetic for the first time in their lives; (*b*) they were feeling more socially competent because of the opportunities provided for social rehabilitation; (*c*) significant people in their lives were encouraged to provide self-enhancing experiences for them.

4.—A close liaison with the parents was considered to be an important factor in helping the children, both academically and socially.

In the judgment of the writer it would be desirable for such special classes as this to be attached to schools which are oriented towards individual progress rather than towards competition in the form of examination positions and streaming by ability. Competition in this form leads to a certain amount of stereotyping which is not conducive to healthy self-evaluation on the part of the children who are unsuccessful within the system. It seems particularly important that children with problems of academic failure and social maladjustment should be placed with teachers who are temperamentally able to accept them and deal with them in a constructive manner.

REFERENCES.

[1] SANDERS, C. (1948). Insecurity and social maladjustment in children. *Brit. J. Educ. Psychol.*, Vol. XVIII, Part III, pp. 148-155.

[2] RAVEN, J. C. (1951). *Controlled Projection for Children.* London: H. K. Lewis and Co., second edition.

[3] STOTT, D. H. and SYKES, E. G. (1956). *Bristol Social-Adjustment Guides Nos. 1 and 2*, U.L.P.

[4] SIEGEL, S. (1956). *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill Book Co.

The Opinions of Young Workers on their own Education in a Secondary Modern School.

By FRANCIS S. JACKSON

(*Abstract of thesis submitted for the degree of M.Ed. in the University of Manchester, April, 1962.*)

PURPOSE OF THE INQUIRY.

The purpose of the inquiry was to obtain at first hand the adult opinions on education of a group of ex-pupils of a secondary modern school and to examine these from the point of view of (*a*) any relevance they might have to the organization and administration of such a school, and (*b*) their bearing upon how the school stands in the esteem of its own ex-pupils, the parents of tomorrow.

THE SAMPLE.

This consisted of 142 young men drawn from four consecutive one-year age-groups. They had left the same secondary modern school at the age of 15 and were, at the time of the inquiry, between 20 and 24 years of age. They represented 83 per cent. of the school's total output of 15-year-old leavers over the four-year period in question, and included a wide range of occupations ranging from undergraduate to labourer. A simple statistical check revealed no significant differences in age and ability-grouping between the 83 per cent. included in the sample and the 17 per cent. omitted.

THE SCHOOL.

The school was a small secondary modern boys' school serving a sea-side town and the adjoining semi-rural area. Its organization and size (two form entry, with a single fourth form) was fairly stable during the period under review. The curriculum incorporated no particularly distinctive features and no unusual teaching methods were in use.

PROCEDURE, AND SCOPE OF THE INQUIRY.

The majority of the respondents (115 out of 142) were interviewed personally by the investigator in their own homes, usually in the evening. Those who, for various reasons, could not be reached in this way completed a questionnaire which covered the same topics as were dealt with in each interview. These were :

(*a*) Respondent's present occupation.
(*b*) Record of further education (if any).
(*c*) Opinions as to the adequacy or otherwise of the school as a preparation for employment.
(*d*) The respondent's views on the relative values of various school subjects.
(*e*) Opinions and ideas on vocational guidance.
(*f*) Free comment or criticism on any seemingly relevant matters.
(*g*) An over-all assessment of each subject's satisfaction with his own secondary education on a seven-point rating scale.

SUMMARY OF CONCLUSIONS.

1.—The majority of the young men in this four-year sample of ex-pupils were favourably disposed towards their own secondary modern school. There was a slight tendency for those in the upper half of the ability range to be more critical (constructively) than their fellows. Extreme dissatisfaction was rare.

2.—There was a distinct association between ability group and strength of demand for higher academic standards. A higher standard of mathematics was particularly requested. The requirements of further education appear to be largely responsible for this.

(3) Almost 72 per cent. of the entire sample (over 94 per cent. of the ablest quarter) had engaged in some form of further education. This had led, in two cases, to the university, and in many others to diploma and certificate courses of various kinds. Such achievements as these, especially in ex-pupils of so small an establishment, need to be remembered by those who would dismiss the secondary modern school as an academic ' dead end.'

4.—Suggested additions to the curriculum (as it then existed) were metalwork, a modern language, and algebra. Requests for metalwork were numerous and appeared to bear no relationship to level of general ability. Demands for algebra, on the other hand, showed a distinct association with level of ability.

5.—Many of the ' second best ' performers felt they had not received their full share of attention, particularly in the basic subjects. They felt they should have been pushed harder, encouraged, and allowed to try some of the work which was evidently considered beyond them.

6.—Music and history came in for heavy criticism as serving no worth-while purpose in the modern secondery school. From the less severe critics came suggestions tnat music might be made optional and that a new approach to history might be made via civics and current affairs.

7.—Mathematics, English, practical subjects, science, and geography were rated by *all* ability groups as the five most generally desirable subjects. From a purely vocational point of view, mathematics, woodwork, and technical drawing were rated higher than any other activities.

8.—Age, as reflected in the Christmas, Easter and Midsummer leaving groups, appeared to be significantly associated with level of ability, the upper ability groups containing a disproportionately high number of the Christmas leavers. Since this phenomenon has been noted elsewhere, its possible effects deserve further study.

9.—Approximately half the respondents declared themselves satisfied with existing vocational guidance arrangements, but some of the more thoughtful suggestions made pointed to the value of adequate school records and to the possibility of the detection of specific abilities as they emerged.

10.—The level to which these young men have risen in their chosen occupations, their general record of stability in employment, and the level to which they have carried their education since leaving the secondary modern school suggest that for the majority of the 142 members of this sample their education at such a school has proved to be a not inadequate preparation for entry into the first phase of adult life.

BOOK REVIEWS

ANSTEY, E. (1961). *Staff Reporting and Staff Development.* London: Royal Institute of Public Administration in collaboration with George Allen and Unwin, Ltd., pp. 96, 12s. 6d.

Applied psychologists who have read publications by Edgar Anstey would expect this book to be meaty, concise and practically helpful. They will not be disappointed. A quarter of the book is made up of appendices showing exact copies of staff reporting forms in current use. In the remaining seventy odd pages Anstey is able to compress, in a highly readable form, balanced accounts of the objectives and methods of staff assessment, of confidential and open systems of reporting, the pros and cons of Promotion Boards and some very worthwhile advice on the design of report forms; he gives us the benefit of his experience in training in reporting and describes three kinds of staff appraisal interviews. A section which particularly impressed me was the five-page critical account of the forced-choice technique. I also liked the examples he gave of the ways in which clever reporting officers could subtly twist the reports so as to secure the promotion of people they liked or the transfer of people they disliked. Obviously, Edgar Anstey has worked with very intelligent people.

On the face of it, this book is a valuable one for Establishment Officers in County Halls and Ministries but not one for educationists. However, those concerned with the grading and assessment of children, of students and of teachers and those who work with follow-up and validation in any field . . . all these will find *Staff Reporting and Staff Development* a very useful handbook. DENIS MCMAHON.

BEREDAY, G. Z. F., and LAUWERYS, J. A. (1962). *The Year Book of Education—The Gifted Child.* Evans Bros., in association with The University of London Institute of Education and Teachers' College, Columbia University, New York, pp. x+541, 63s.

Readers of this *Journal* will probably find this volume of the Year Book one of the most valuable of the series for it deals with theoretical and practical problems of giftedness to which society is urgently demanding answers of psychologists. The three major sections are headed *Psychological Theories, Detection of Ability and Selection for Educational Purposes* and *Increasing the Pool of Talent.* These are preceded by an introduction by Sir Cyril Burt which must surely be one of the most distinguished pieces of writing in educational psychology in recent years. Here in under sixty pages he broadly surveys the whole field and so orientates the reader for the detail which follows and provides him with a critical evaluation in advance of much of what appears in the rest of the book.

Section One deals with theories as to the nature of giftedness and the educational practices likely to stem from them. Of particular interest are Ruth Strang's introduction, Guilford's *Parameters and Categories of Talent* and Torrance's *Measurement and Development of Creative Thinking Abilities.* The methods of selection of gifted children in a wide range of countries which are described in the Second Section are on the whole retrospective studies reflecting existing techniques which have grown up in the countries of their origin with little reference to advancing psychological knowledge. Most are based upon examinations and frequently include tests of academic aptitude. Some appear unsophisticated to British eyes. There are, however, some valuable accounts of researches designed to assess the efficacy of procedures, as, for example, in Passow's survey and Husén's work in Sweden into the age at which selection should take place. Contrary to many beliefs Husén concludes that early selection of the academically gifted does not benefit either those selected or those left behind.

The third section concerns attempts of modern technological communities to meet their insatiable demands for people of high talents, and of rapidly developing countries to increase their available human resources as quickly as possible. The proposed solutions are the familiar ones of recognition followed by appropriate educational environments designed to encourage maximum development commensurate with individual genetic limitations. There is less than one might have hoped of discussion of the social problems satirised by Michael Young which inevitably arise as segregation of the most able becomes more efficient, though Anderson does touch on them in his chapter on *Dilemmas of Talent Centred Educational Programmes.*

The levels of writing are inevitably uneven and few but a reviewer will have the fortitude to read the whole, but there is no doubt of the book's value for it gathers together a great range of thinking on the topic of giftedness and provides a very complete entry into the many aspects of the field. L. B. Birch.

Cleugh, M. F. (1962). *Educating Older People.* Tavistock Publications, pp. 185, 28s.

This interesting book is an attempt to generalise the experience of the author in tutoring students for the diploma in teaching backward children working at the Institute of Education, University of London. The author makes considerable use of what these students have said themselves in reply to queries about their experience of the course. The book has all the virtues and defects which arise from the use of such a method: flashes of insight occasionally light up the general picture of banality and do-it-yourself psychoanalytic rumination which constitutes the second half of the book. The treatment is very uneven since Dr. Cleugh seems to work at two different levels, as a rational analyst of that experience which she shares with trained teachers (the most fruitful part of her communication) and as an amateur case-worker who works on the most superficial and subjective level of analysis of ordinary human contact. In the days of my youth I became familiar with the kind of content which is now being widely publicised as the latest advances of psychological insight—except that in those days it was described, perhaps unkindly, as "stair-heed gossip." I quote:

"*January* 29.

> Tutor: Are you keeping your head above water?
> Mrs. S.: Yes, much better this term.
> Tutor: Good. Take it easy.
> Mrs. S.: It's only *just* above water, but it is above. (*Seems steadier and more relaxed, in spite of brick on 17th.*)." p. 129.

The first hundred pages deal in a systematic way with the problems of educating older people, under the chapter-heads of *Getting things done, lecturing, discussion, supervision of practical work, examinations.* This part of the book might be regarded as a handbook for adult tutors: it is helpful, informed and systematic. Although it is on an *ad hominem* level with very little reference to the scientific studies which have been made of the adult learner, there is nothing which one would fault as contrary to one's experience or conclusions about the adult.

The second half seems to have been the result of some kind of pressure to produce a "proper book" in the usual Tavistock format and style rather than a mere brochure of "suggestions for the use of tutors." Dr. Cleugh expresses the view that although the tutor is not a professional therapist he can (and *ought* to?) serve a therapeutic purpose. It is the attempt to become self-conscious and centrally concerned about this function with which one would quarrel.

However, there have been so few books published in this country dealing with the adult student that one must compliment the enterprise of the author and the publisher, and hope that the book will be widely read since it will certainly be of value to many different kinds of reader. Although not the most economical in terms of time, one occasionally is illuminated in listening even to "stair-heed gossip."

John McLeish.

GORDON, I. J. (1962). *Human Development.* Harper, pp. 400, 46s.

This book traces the development of the child from conception to manhood, preserves a good balance between the forces of heredity and environment, and successfully combines the physical, psychological and sociological aspects. The adjective which best describes the first two-thirds of the book is ' sound,' and this is by no means damning with faint praise. Two late chapters, on " Exceptional Development " and " The Continuing search for Identity " are very good. In general, both effectiveness and interest tend to increase as one reads on.

The author endeavours to study the developing child as a whole while also examining the forces bearing on him at certain times in his life, and seeks to adopt a ' unifying theoretical position.' The development of the self ' is a transactional process between the developing organism and its changing social environment.' He tries, to combine an " external " approach of looking at the child's home, his peer group, his school and his society, with the " internal " approach of looking at how he perceives these various situational forces, and how he perceives his own body and his own growth.' The theoretical account which is constructed on this basis is well supported by concise but informative statements of recent research. Though there are some references to British work, there could profitably have been more.

Though the book is particularly sound there are a few statements which need careful examination—though one indeed has seen them elsewhere. Gordon adumbrates that " The direction of (man's) behaviour is always, in normal development, towards increased complexity." The intention here is good, but one wonders about the word ' always.' When man, for example, reduces a chaos of isolated facts into order by classification is he not simplifying and consolidating the position, maybe preparing for further complexity but not increasing it for the moment ? Similarly, we read " the increase in complexity and organisation of the person continues throughout his life." Does this statement include old age ? Much more important, however, is the widely accepted theory of ' equilibrium ' which Gordon states once again. The human organism " tries to maintain a steady state, i.e., it attempts to maintain itself and to balance the forces within itself." While there is a central truth here the words do not describe exactly what is happening, and Gordon (and many other writers) contradict themselves when they maintain that the organism "at the same time develops towards a new steady state." The reviewer's position is that the developing child must contain, if healthy, a tendency towards greater maturity, towards development, therefore *ipso facto* it cannot be in a ' steady state.' Physiologically, it is obvious that the child contains within himself the tendency to change, and this is also true psychologically. This often repeated concept of ' equilibrium ' needs a much deeper examination and a much more thoughtful comment.

R. R. DALE.

HADFIELD, J. A. (1961). *Childhood and Adolescence.* Pelican Books, pp. 286, 5s.

There are few books of child psychology, covering development from infancy to adolescence, which can be recommended confidently to parents who lack specialist knowledge of psychology. Dr. Hadfield has provided such a book by writing for " ordinary intelligent parents of ordinary intelligent children." He has drawn on his experience as a psychiatrist to give many examples of children's behaviour and dfficulties, and to make recommendations to parents about the treatment of children. These are founded on his belief that the fundamental need for normal development is a sense of security. It follows that he gives little emphasis to sexual problems as primary causes of neurotic symptoms ; he considers that mental health in childhood depends on parental love, on discipline which combines kindness with firmness, and on freedom from the inhibiting effects of unnecessary restrictions or ' over-protection.' Some readers may feel that genetic constitution and social environment also play an important part but, in so far as Dr. Hadfield considers the latter, he finds that a sense of inferiority and competitiveness play minor roles in the development of most children.

Following an introductory discussion of parents' problems in bringing up children, chapters two to four describe the stages of maturation, phases in early development and the organisation of personality during childhood. In chapter five emotional development in adolescence is described at some length. More serious neurotic disorders are dealt with briefly in chapter six and, in a last chapter, several 'golden rules for parenthood' are suggested.

Dr. Hadfield stresses the importance of allowing children to live fully through each phase of development, expressing each impulse before it is disciplined and directed. In disciplining children, he recommends the use of encouragement with a minimum of blame; but he considers that punishment of wrong-doing should be inevitable, with the proviso that the child's motives should first be understood.

The long chapter on adolescence is likely to be most appreciated by parents; for whereas writers in some popular journals deal knowledgeably with the difficulties of young children none deals so soundly with the problems of adolescence. The four stages of sexual development outlined here may be a new conception to readers already having some knowledge of psychology.

This is a well-written book, in a pleasant and persuasive style with a fund of illustrative anecdotes; the views expressed are likely to impress most parents as eminently sensible. It should do much to help them, as well as stuents and teachers, to understand the development of children and to alleviate their difficulties.

RUTH BEARD.

HAWKES, GLENN R., and PEASE, DEMARIS (1962). *Behavior and Development from* 5 *to* 12. New York: Harper and Bros., pp. ix+375, 46s.

This is rather a charming book which one would wish to have on one's shelves (rather than order for the Library). It has an attractive format, clear print and good photographs. The content is remarkably free from distortion, considering the brevity of treatment of a wide variety of topics. These are presented in fifteen chapters, each with headed sections and sub-divisions. The integrity of approach, and the deft, direct style, impose their own unity, and can stem only from being thoroughly at home with the subject-matter, and having masterly powers of selection. The book is at once easy to read and motivating to more advanced study, for which references are given at the end of chapters. The work of British writers is included (e.g., for Ch. 9, "growth and function of intelligence" mention is made of research by C. Burt, M. Dunsdon and J. Fraser-Roberts, and G. Seth).

The authors state that their book is about children in their natural habitat—their homes, their neighbourhood and their schools. The reader who expects to have a clearly separated view of each year from 5 to 12 will not find this, except for a very condensed "summary of age characteristics" (agreeing with Gesell's accounts) in Ch. 14 (pp. 315-330). The emphasis is more on the 5–12 'school-age,' as a whole, except for incidental differentiation.

The book is ideal for groups of the weekly tutorial or seminar kind, either in training colleges, adult evening classes, or in university introductory courses in child development with a social psychology orientation. At the end of most chapters is a list of "Study Problems." These are thought-provoking practical exercises not too onerous to be carried out with enjoyment in limited time. There is a chapter (13) which gives details and principles for observing and recording children's behaviour. Appendix A and Appendix B list, respectively, films and selected pamphlets, each with a descriptive sentence. The last chapter is "Adolescence, a look ahead" with suggested readings, which would give a study group the basis for a follow-up course.

It is assumed throughout that the parents and teachers for whom the book is mainly written will wish to use the knowledge provided. There is also a specific chapter (12) on guidance, with a word of caution in the preface: "The understanding parent or teacher will refrain from the use of specific techniques unless he has a philosophical understanding of the child and the demands being made upon him."

A lively feature is the introduction of quoted extracts, sometimes from children themselves, e.g., exchanging jokes and riddles (p. 227), or an 11-year-old writing about teachers (p. 53) : " There are some teachers who have just gotten tired of children because of the bothersome things they do, and do not like any of them any more. This kind do not help you learn very much . . . " The authors of this book are clearly not in this category. " Being around children," they say (p. viii) " may make the hair gray, but it keeps the heart young." It is pleasant to be able to recommend a book of high academic standards written with enthusiasm in a manner which meets the needs of the less academic and of those students and professional workers who would benefit from a moderately short, gay but not superficial text for this relatively little studied 5–12 age.

L. R. BOWYER.

MANNHEIM, KARL, and STEWART, W. A. C. (1962). *An Introduction to the Sociology of Education*. London : Routledge and Kegan Paul, pp. xvii+187, 21s.

This work is chiefly of historical interest. While it may be the last of Mannheim, no one would claim that it was the best. Its fifteen chapters, arranged in five sections, are uneven in value and loosely strung together. It purports in its title and Introduction to constitute an introduction to the sociology of education. It does not. No beginner who read this book would, in fact, have been introduced to the major pre-occupations, problems and achievements, still less the methods, which characterize educational sociology in the mid-twentieth century.

One-third of the book is devoted to psychological topics : resumés of the more obvious and unsophisticated theories of learning and personality development. There is some advance beyond associationist theory : the laws of recency, frequency and effect are rehearsed, but attention is given to the social context in which learning takes place, to the assumption of roles within social institutions. This is both too little and too much. It is doubtful whether the beginner would make much of a passing reference to Allport's theory of ' functional autonomy ' ; for others it is really rather old hat. The chapters on social influences on personality development draw heavily on G. H. Mead ; they are written with only a bare hint of the contribution which the cross-cultural studies of social anthropologists have made to this field of inquiry.

The first third of the book and the last deal with matters which are more properly sociological. It is interesting, however, that in a chapter which usefully discusses teaching, training, and instruction, the temptation to explore the sociological implications of these distinctions is firmly resisted. The sociological chapters contain much that is platitudinous and trite. An ambitious survey of the ideals of the Graeco-Roman, Hebrew, Medieval, Renaissance, and post-Renaissance worlds illustrates the point that different ages modify the ideals (for example of ' the gentleman ') which they inherit, and that ' current attitudes have historical antecedents.'

Nowhere in the book are we introduced to the sociology of education as an experimental science concerned with the rigorous empirical testing of hypotheses. In the final chapter, it is true, a number of suggestions are made for research : a ' survey of the existing educational institutions in this country and their interrelations . . . ' ; the factors in modern life (work and leisure) which make for the disintegration of the personality ; the proper recruitment and role of élites if the cultural decline of our society is to be halted ; ' the balance between the contributions of the home, church, school, factory, and public life towards moulding individual character.' There is little apparent awareness of the formidable technical problems involved in setting up inquiries which might yield answers to these great questions. This reviewer remains unconvinced that " the right command of primary sources and of the existing literature on the subject already offers answers to a great number of them."

F. MUSGROVE.

MONEY, J. (Edit.) (1962). *Reading Disability. Progress and Research Needs in Dyslexia.* Baltimore: Johns Hopkins Press, pp. x+222. No price given.

The papers published here were submitted at a symposium held in the Johns Hopkins Medical Institutions in 1961 under the auspices of the Association for the Aid of Crippled Children. There are thirteen symposiasts drawn from the general fields of education, psychology and medicine. In itself, this is a reassuring sign for medicine and paedagogy have been too slow in coming together in joint studies of mutual interest.

Workers in the field of severe reading disability will not be surprised that no dramatic break-through is reported or that Howes writes " the field is still at a stage where we need to find out exactly what dyslexia is as well as why it occurs." It is noteworthy, however, that in spite of the opinions given by such as Torpe and Malmquist (who were not at the Conference), that no doubt is expressed as to the existence of the syndrome of specific developmental dyslexia. It is seen as a dysfunction of written language and communication resembling the sequelae of some brain injuries.

Most of the familiar topics are dealt with and brought up to date by reference to the most recent work and there are some topics which will be unfamiliar. On the educational side there would seem to be little new to report with regard to phenomenology, or to recognition and treatment of the condition. As Schiffman indicates we still await answers to the questions : What is the aetiology ? How can we make an early diagnosis ? What are the best remedial procedures ? It seems clear that answers are unlikely to arise from further large scale surveys and from correlational and factorial studies. It is pointed out by Zangwill that detailed clinical and experiemntal studies are most likely to throw light upon them. The latter half of the book is concerned with such studies though the papers are largely statements of points of view of workers in particular fields rather than the reporting of experiments.

Birch, in his paper on *Dyslexia and the Maturation of Visual Function*, insists that dyslexia is not itself a problem capable of being investigated since it may derive from a variety of aetiologies. This, he suggests, makes it necessary for any worker to decide at the outset which aspect of the problem he wishes to explore ; but, having made a decision, this does not mean that he regards other aspects as unimportant. It is the main value of this book that workers who have made decisions of this kind have placed their considered opinions side by side so that each can be correlated with all others and the way forward seen a little less obscurely.

This is essential reading for all workers in this field. It follows on and supplements M. D. Vernon's (1957) definitive volume *Backwardness in Reading* bringing the bibliography up to 1961.

L. B. BIRCH.

SCOTT, W. A., and WERTHEIMER, M. (1962). *Introduction to Psychological Research.* New York : Wiley, 60s.

The writers of this book have set out to provide help to those who are embarking upon psychological research for the first time. They use the word psychological broadly and what they have to say is relevant to all the social sciences.

There are some to whom such a book will be an anathema. They would aver that the only way to learn how to do research is to embark upon it and that without such a process the knowledge acquired by the theoretical study of research procedures ls ephemeral and jejune. But this view forgets the position in which the newcomer finds himself. It ignores the fact that psychological research techniques are highly sophisticated, are not so easily acquired by contagion as in many other fields and exist in considerable variety. It does not, moreover, take cognisance of the fact that, for many new research workers, it is the existence of techniques as well as of problems which often provides orientation and a conceptual framework for their proposed studies.

The need for the guidance which such a book can provide would almost certainly be granted by those who have numbers of higher degree students passing through their hands, yet the attempts to provide it have been rather few. Of these, the present is the best which has yet appeared. It deals in a co-ordinated manner with the things one would expect, with the processes of planning and design in fair detail, with the techniques of instrumentation, verbal, observational and experimental, with sampling and with analysis, interpretation and reporting of results. Throughout, the hand of experience is clearly revealed and the discussion is informed, up to date and most clearly laid out. Yet, even so, if the purpose of the book was to inform and guide the beginner it might still have been better. It tends, as do all books in this field, towards a discussion of what research is about, as if it were a discipline in its own right, rather than concerning itself with how to do research, insofar as it is ever possible to describe this. This is, of course, a simplification and generalisation which in some detail is unfair to the authors. To several of their chapters they have added supplements which deal with the ' how ' rather than the ' why ' and these chapters, and not only their supplements, are among the best in the book. Yet, too much of the 400 pages of text is taken up with discussion which is excellent for the relatively sophisticated but which is inappropriate to the person seeking fairly specific guidance during his novitiate. One suspects at times that the writers wanted to present a straightforward yet informed guide book to the neophyte but were deterred from writing the whole book in this way by the fear of being thought unscholarly. If this were so, it was an unnecessary fear both from the viewpoint of the subject matter and from the standing of the authors themselves. Had they written more specifically a guide book in the way indicated, they would surely have found space to deal, *inter alia*, with questionnaires, to which so many beginners are attracted, with a treatment of non-parametric techniques in sufficient detail to give the newcomer some orientation and, surprising omission, with factor analysis in outline and with the types of study amenable to it. If space was short they could well have omitted the twenty-five pages of standard tables.

But these observations, notwithstanding, the book remains the best of its type and should certainly be in all libraries, in the hands of those sophisticated enough to skip some parts and to benefit from others and with anybody else who is not deterred by the rather high price ticket which American books seem to attract in this country.

G. E. R. BURROUGHS.

VENESS, THELMA (1962). *School Leavers.* London : Methuen and Co., Ltd., pp. xxv+252, 25s.

This book is concerned with young people's outlook, chiefly on their own personal future. What do they expect from life in its two main spheres, family and work ? In particular, and this is the problem which first stimulated the inquiry, what are their ambitions ?

The study was carried out with 1,300 boys and girls in two counties, one in the West of England, and the other in the vicinity of London. They were in grammar, technical and modern schools, with ages ranging from 13-plus to 17-plus, depending on leaving age. The methods employed were, to borrow a phrase from Wilkins, of the ' grape shot ' variety : questionnaires, interviews, paired comparisons and so on. The underlying assumption, shared by the reviewer, is that in certain kinds of research a multitude of relatively coarse approaches is more likely to yield a rounded picture than a single approach, however thorough. The key technique on which the author relied most heavily—though without omitting frequent checks from different sources—is the ' restrospective autobiography.' The writer was asked to imagine himself at the end of his life, and review it from that perspective. This technique, which was previously used by Stoetzel in his study of Japanese youth, produced a rich harvest of fascinating material.

Some aspects of ideas about marriage and the family may be cited by way of illustration. Current trends are reflected in the favourite age for marrying, between 20 and 25. Although it is not directly stated, one might infer from the quotations a remarkable absence of the romantic notion of marriage (remarkable because the prevalent stress in the mass media) : " This job I stayed at until I got married at the age of 20. I had a good husband and we never quarrelled." It is mostly rather matter-of-fact. This also applied to the references to children ; there was little to support the view that frustrated aspirations tend to be projected on to fantasy children, and scant emphasis on education as a goal. Very few mentioned separation or divorce, but a curious feature was the anticipated high mortality rate of spouses.

Attitudes about work, knowledge of jobs and job choices are extensively dealt with, and this section of the book wi.ll be found most useful by those concerned with the practical problems of school leavers. The author introduced an interesting analytical categorization of job-choice orientations in terms of ' tradition-,' ' inner- ' or ' other-directed,' which deserves further development.

The findings concerned with ambitions have aroused most widespread interest and led to the vociferous airing of prejudices. A common reaction to the book was that it reveals the youth of today as dull and unenterprising ; had the results been in the opposite direction, the line would no doubt have been that modern youth is aiming too high, with inevitable later frustration. Here it will only be noted that the statistical analysis of ' ambitiousness ' items has as its outcome a scale on which technical school boys rated highest. The problem is of course,, a difficult one, and the author herself did have some doubts on this score. It is, perhaps, questionable whether any single scale can satisfactorily measure ' ambitiousness ' across the whole range from unskilled to higher professional work. Another factor, which should be kept in mind with regard to all the comparisons made between schools, is that a majority of the grammar school pupils investigated were somewhat older than the rest.

These comments are only possible because extensive and detailed tables are provided. Readers tempted by the lively style to skip perusal of the tables should resist the temptation—they contain much of the meat. Altogether the book has a great deal to offer to everybody concerned with young people, but perhaps most of all to educationists (a sample of their products) and to youth employment officers (a warning that what they are apt to regard as raw material is already partly processed.).

G. JAHODA.

VYGOTSKY, L. S. (1962). *Thought and Language.* M.I.T. Press and Wyley.

This is an English translation of a book written getting on for thirty years ago which appeared again in Russian only as recently as 1956. It shows remarkable insights into the nature of thinking, and, in its essentially functionalist approach, is an appropriate work to put before the English reader today.

After presenting his problem, which is the growth of thought as shown in the construction of concepts, Vygotsky then discusses Piaget's and Stern's views of language development. The comment upon Piaget is, of course, based only on his earlier works and the publishers include with the book a separate paper of some thirteen pages, setting out comments made by Piaget in 1962. Vygotsky is critical of the extent to which Piaget's theory leans on the concept of maturation. However, Vygotsky's own contributions form the really important parts of the book, particularly as set out in Chapters 5, 6 and 7. In Chapter 5 he describes how conceptual grouping by children grows from heaps having no objective basis, through complexes with no logical basis but possessing some of the qualities of concepts, finally to true concepts. In the next chapter, on the development of scientific concepts, he has much to say to the educationalists, in particular contrasting spontaneous with non-spontaneous concepts. The latter he exemplifies from school work in science and modern language learning, and shows how the two types of concept tend to come together in the course of learning. Finally, there is a thoughtful discussion of thought and word, in which the growth of speech and thinking are considered together and in which Piaget's views on the socialising of egoentric speech is commented upon and checked and criticised by empirical study.

In all the book is a ' must ' for workers in the field of the growth of children's thinking and language. Those who know the publication by Piaget and Inhelder, *La genese des structures logiques elementaires*, will also recognise how similar is the picture presented by Vygotsky of the growth of concept formation to that set out by the Geneva School in their publication. The book is, however, more than a psychological treatise. Much of it offers a method of education, and so should be read by research workers in this field.

E. A. PEEL.

WISEMAN, S. (Edit. 1961). *Examinations and English Education*. Manchester : Manchester University Press, pp. xx+188, 21s. net.

In 1959, the subject chosen by the School of Education in Manchester University for its annual public lectures was the problem of examination. There were, we are told, several reasons for this choice—the growing interest of secondary modern schools in examinations of an external type, the feeling that the time was ripe for taking stock of post-war changes in examinatious which affect either primary or secondary schools, and finally, the recent developments in the techniques of examining and the need for relating them to the purposes the examination seeks to fulfil. The present volume, edited by Professor Wiseman, is an expanded version of the lectures delivered on these topics, with an additional chapter from Professor Oliver dealing more especially with the recommendations of the Crowther Report.

After a short introduction, the series opens with a ' historian's view of examinations ' by Mr. Norman Morris. This is followed by lectures on examinations in relation to the grammar school, the modern school, and the primary school, respectively, by Dr. Petch (Secretary to the North Universities Joint Matriculation Board), Mr. Lucia (formerly Secretary to the Associated Examination Board), and Dr. Wiseman. Dr. Wiseman also summarizes the methods and results of assessing the efficiency of examinations ; and Professor Oliver discusses ' examining general education,' and then winds up the course with a paper on the problems of education and selection.

All five contributors are conscious that both selection and examinations have their limitations and their drawbacks, and therefore, call not only for some defence if they are to be retained, but also for certain improvements if current criticisms are to be met. They are, however, equally convinced that in recent years criticism has been pressed too far. At bottom, as Dr. Wiseman observes, the objections to selection, and particularly to selections by means of intelligence tests, are " social and political in origin rather than educational and psychological." But the familiar arguments have been strongly supported by sociological writers, like Dr. Floud and Dr. Halsey. In their book on *Social Class and Educational Opportunity* they assure us that " measured intelligence is well known to be largely an acquired characteristic," and therefore unfair to children of the " unprivileged classes." But this, says Dr. Wiseman (quoting Dr. Warburton), is rather like declaring that " the English Channel is well known to be wider than the Atlantic Ocean." No one, he adds, who has really studied recent research on the subject " can possibly deny the very substantial genetic component in tested intelligence " ; and, as actual records show, the introduction of such tests has undeniably helped to do fuller justice to those who are socially or culturally handicapped. " The intelligence test *must* be used if we are to get anywhere near equality of opportunity ; and we would be more successful in reaching this objective if the test had greater weight."

Professor Oliver defends much the same point of view. From Plato onwards, so he reminds us, philosophers have maintained that educational systems have a dual function—to train and to select ; and there is the inevitable temptation to over-emphasize one function to the neglect or impairment of the other. Current discussion, too, has tended to forget that efficient selection implies and facilitates social mobility. Accordingly, in the interest of the community as well as of the individual, it is to be hoped (he says) that ill-informed attacks on examinations and selection tests will be stoutly resisted. The Crowther Report " tries to escape the necessity of selection by reviving the supposed distinction between selection and qualification." The G.C.E. has come to serve both purposes; but the Report proposes to revert to a separate examination for each. Professor Oliver holds that the distinction is unrealistic: if the number qualified exceeds the number of places, we must select; and we ought, so far as we can, to select the best. Hence, he argues, the Crowther Committee " would have done better to address itself to the improvement of the G.C.E.—the only public examination at this level which has achieved national currency."

How such improvements might be achieved is admirably outlined by Dr. Wiseman. Hitherto, he believes, " educational psychologists have tended to ignore the wider educational implications of their work, while the teacher-examiners of the traditional boards do just the reverse ": they overlook the need and the possibility of increasing the sheer efficiency of their techniques. In the United States, " comparable boards have research sections engaged full-time on such problems "; and he rightly urges that there is pressing need for similar studies over here. He sums up the whole book by saying that "good examinations are both useful and desirable—without them education would be much less effective than it is: let us, therefore, make sure that the examinations we do use are efficient in the fullest sense."

CYRIL BURT.

WILSON, HARRIET (1962). *Delinquency and Child Neglect.* George Allen and Unwin, pp. 195, 25s.

There have been numerous studies of the relationship between family background and juvenile delinquency, but few of these have been studies in depth of the ' problem family ' in which the children suffer some measure of neglect. Dr. Harriet Wilson's thoughtful contribution to our knowledge in this field is, therefore, to be welcomed and merits careful study. Basing her book on an enquiry into a group of problem families in an industrial city, she sets out to define what is meant by the ' problem family ' and to discover to what extent such families are the breeding ground of delinquency.

After discussing recent trends in studies of juvenile delinquency (she surprisingly omits reference to some important work of relevance, for example that by Stott, in this country, and Hewitt and Jenkins in the United States) and making some interesting comparisons between British and American culture-patterns, Dr. Wilson goes on to describe the patterns of living of fifty-two problem families, selected as her sample on the basis of performance-inadequacy, under the headings of solvency, health and education. She finds that the most disabling feature of the problem family is its social isolation, while emphasizing that these families are not, as is commonly assumed, a homogeneous entity. Delinquency is part of the pattern of juvenile behaviour among families exhibiting signs of child neglect, the main factors dominating the children's lives being poverty and weak relationships with their parents and other adults. The chief characteristic of the juvenile delinquent from the problem family is his complete unawareness of the moral significance of his actions.

The book is refreshingly free from sociological jargon, and tries to avoid unwarranted generalisations. Dr. Wilson has kept an open mind throughout her enquiry and is, on the whole, careful about methodological considerations, though the children have been studied less systematically than the parents. For example, a number of conclusions about the level of the children's intelligence are drawn from data which are incomplete and, being partly based on group tests given at the infant school stage, hardly reliable; and the author's reason for obtaining teachers' behavioural assessments in an unstandardized form seems inadequate. Furthermore, can one justifiably infer the presence or absence of maladjustment from the correlation between I.Q., attendance, and educational attainment (p. 73)?

A slight re-arrangement of the material might have made the main findings and conclusions stand out more clearly from the mass of detail ; and some of the tables are unnecessarily obscure. There is, however, likely to be little disagreement with Dr. Wilson's main conclusions and with her practical proposals in the final chapter, where she advocates, as preventive action, measures which would help the father with a low wage and large family (so that he would no longer be financially little better off working than if unemployed) as well as aid those families permanently dependent on assistance from the state ; realistic family planning methods ; an increase in the number of nursery schools ; and the provision by local authorities of specialized supportive help of a personal kind. MAURICE CHAZAN.

YATES, AUBREY J. (1962). *Frustration and Conflict.* London : Methuen, pp. x+236. 28s.

The author states that this book is not intended to be " an original contribution with respect to theory," but rather " a systematic examination of the accumulated knowledge in the areas of frustration and conflict." These topics have been extensively investigated by psychologists, both clinically and, in more recent years, experimentally. Dr. Yates is concerned primarily with work of the latter kind, but a most useful bibliography of well over three-hundred items shows how widely he has, nevertheless, cast his net.

The book comprises a brief introduction and five main chapters. The first is a very thorough review and critical analysis of the work of Maier and his colleagues on frustration and fixation. The author frequently comments on the neglect of Maier's findings by other psychologists and remarks : " There is no doubt that the lasting value of his work has been obscured by sheer prejudice and *a priori* convictions." He, nevertheless, admits that Maier's results and theoretical interpretations are still difficult to reconcile with other contemporary approaches. The second chapter provides a detailed examination of the frustration-aggression hypothesis of Dollard, Miller, *et al.* and of the very considerable number of experiments which this has provoked. The author's verdict is that work in this field has proved " extremely profitable." Next comes an account of studies on frustration and regression. It includes an exhaustive critique of the famous experiment by Barker, Dembo and Lewin. Dr. Yates notes how extraordinarily few attempts have been made to repeat their findings, a need which is surely pressing, since the original enquiry was conducted without a control group and the statistical treatment of data was often unsatisfactory. The fourth chapter is on conflict. Short sections are devoted to the work of Luria, Lewin and Cattell, but the main analysis is concentrated on the extensive work of Miller and his colleagues. Some most ingenious experiments are discussed in this section, including a particularly brilliant *tour de force*, by Murray and Berkun.

The final chapter is entitled " Frustration, Conflict, Learning Theory and Personality." Here Dr. Yates tries to draw the strands together, and at once the influence of Hullian learning theory is apparent. Recent efforts to incorporate ' frustration ' into learning theory as an intervening variable are described, as are many experiments relating to these attempts. Eysenck's theory of personality, itself of course, very considerably influenced by Hullian theory, is also favoured, and hypotheses are put forward linking frustration and conflict with neuroticism and introversion-extraversion.

As a whole, this book is to be highly recommended. Whatever one may feel about the author's theoretical leanings, serious attempts to test rigorously theories about frustration and conflict are of very recent origin and should be encouraged. Dr. Yates' foremost contribution has been to provide a careful, objective and wide-ranging digest of this interesting, and often exciting, current research. For the educationalist, a useful feature is his detailed consideration of the effects of punishment, which are shown to vary greatly under different conditions. More generally, this book provides an essential basic reference both for those contemplating serious research into frustration and conflict and for all advanced students of these topics.

G. W. PILKINGTON.

Some Important Books

THE DEVELOPMENT OF THE INFANT AND YOUNG CHILD

By Professor R. S. Illingworth, M.D., F.R.C.P., D.P.H., D.C.H.
324 pages, 95 illustrations. 27s. 6d.

THE NATURAL DEVELOPMENT OF THE CHILD

Fourth Edition. By Agatha H. Bowley, Ph.D., F.B.Psy.S.
224 pages, 84 illustrations. 10s. 6d.

THE PSYCHOLOGICAL CARE OF THE CHILD IN HOSPITAL

By Agatha H. Bowley.
48 pages, 8 illustrations. 4s. 6d.

CEREBRAL PALSY IN CHILDHOOD AND ADOLESCENCE

A Medical, Psychological and Social Study.
Edited by Professor J. L. Henderson, M.D., F.R.C.P.(Ed.). 411 pages. 35s.

E. & S. LIVINGSTONE, LTD.
TEVIOT PLACE, EDINBURGH

THE
BIRMINGHAM
PRINTERS LTD
PRINTERS & BOOKBINDERS

High Class
Technical
Publications

42-44 HILL STREET
BIRMINGHAM 5
TELEPHONE: MIDLAND 0196-7

Educating Exceptional Children

SAMUEL A. KIRK

Director, Institute for Research on Exceptional Children, University of Illinois.

Various kinds and degrees of exceptionality are considered—mental, physical, and those caused by emotional disturbances. The conclusion arrived at by this leading authority will be of the utmost importance to educational psychologists, to teachers in special schools, and to students in training.

432 *pages*, 28 *graphs and figures*. 37*s*. 6*d*.

Speech Difficulties in Childhood

A Practical Guide for Parents and Teachers

RONA M. WILLIAMS *Licentiate of the College of Speech Therapists*

" Lucidly, simply, and yet without any loss of accuracy, Miss Williams describes the whole range of speech difficulties in childhood and very persuasively calls on teacher and parent to play their parts in helping children to overcome their disabilities. There are few teachers whose range of professional understanding will not be broadened by reading this book." PROFESSOR M. M. LEWIS in *The Teacher*.

Illustrated. 15*s. net*.

HARRAP *182 High Holborn, London, W.C.1*

VOLUME XXXIII JUNE, 1963 PART 2

THE REDUCTION OF SOCIOECONOMIC BIAS IN INTELLIGENCE TESTING

BY R. S. MACARTHUR AND W. B. ELLEY

(*University of Alberta*)

SUMMARY. This study was designed to investigate the extent of socio-economic bias in selected intelligence tests, with a view to identifying and constructing instruments which might better estimate the intellectual potential of pupils from deprived cultural and educational backgrounds. For such pupils the conventional intelligence test is unsuitable as a predictor of later educational progress if adaptive treatment reduces the effects of their educational handicaps.

Nine promising culture-reduced tests and subtests were administered to a representative Edmonton sample of 271 twelve and thirteen-year-old children, and the results analysed in relation to measures of socio-economic status, verbal intelligence, achievement and school marks. City-wide testing surveys conducted during the previous four years provided data on the same children, making longitudinal comparisons possible. The results supported the hypothesis that it is possible to measure a broad component of intellectual ability with significantly less cultural bias than is found in the conventional intelligence test. Raven's Progressive Matrices proved the most useful test in the battery since it showed high g loading, consistent and minimal relation with socio-economic status, no evidence of cultural bias by items, and moderate correlations with school marks. The Cattell test of g and the Lorge-Thorndike Figure Analogies subtest showed similar promise, since they met most of the criteria of the study in shorter testing time.

I.—INTRODUCTION

INTELLIGENCE tests are frequently evaluated by mechanical investigation of their predictive validity in relation to academic criteria. Usually underlying such an empirical approach is the assumption that only one or two pre-arranged and rather standard treatments will occur between test and criterion judgments.

Cronbach and Gleser (1957) have drawn attention to a distinction between fixed-treatment and adaptive-treatment situations in decision-making based on the results of mental tests. In the former case, treatment is specified *a priori*, and maximum utility is related to predictive validity coefficients; individuals are selected or placed simply in accordance with their promise in relation to a predetermined criterion with predetermined treatments, and little regard for the qualities underlying the test is necessary. In educational prediction, the intervening treatments are usually limited by a relatively fixed curriculum and relatively standard teaching techniques.

In adaptive teaching decisions, however, atypical individuals may be assigned quite different treatments according to their present status and indicated potential. In an educational setting, immigrant children may be given supplementary English coaching, children from deprived backgrounds may benefit from remedial reading, and sensory-handicapped children may be offered special aids. Educational philosophies which emphasize the development of potential to its maximum require serious consideration of adaptive rather than fixed treatments, and to aid in such adaptive treatment, the use of mental tests which can provide estimates of potential ability, divorced as much as possible from inequalities in cultural background. The conventional verbal test of intelligence, designed as it is for subjects raised within a specific cultural range, has limited suitability for identifying potentially gifted children from deprived or foreign backgrounds, or with restricted formal education. The writers of this article suggest that various non-verbal tests for use in adaptive treatment settings have received insufficient research attention.

It is further suggested that the choosing of a test to help identify potential for attainment of distant and general criteria, assuming the possibility of intervening treatments appropriately adapted to realize that potential, can be facilitated through the 'construct validity' approach advocated by Cronbach and Meehl (1955). The evaluation of tests of educability should place more emphasis than is customary on theoretical analysis of their relation to the concept of intelligence as developed by experience, with proportionately less emphasis on their 'cold-blooded' predictive validity under limited fixed treatments. If potentially suitable tests can be examined in the light of hypotheses generated from an adequate theory of intelligence, some progress may be anticipated in the development of culture-reduced tests suitable for classifying and selecting children from diverse cultural backgrounds.

It was the purpose of the present investigation to identify tests or subtests which, for a sample of 12 and 13-year-old pupils from a variety of socioeconomic backgrounds in Edmonton, Canada, could on the one hand, be regarded as valid measures of intellectual ability, while on the other hand, could demonstrate a minimum of socioeconomic bias. Parallel studies with children from different ethnic and language backgrounds are now in progress (MacArthur, 1962).

II.—PREVIOUS RESEARCH.

Binet himself realized that his measuring scales were sensitive to 'scholastic and extra-scholastic acquisitions' as well as 'intelligence pure and simple' (1916). Further analyses by Stern (1914) and Burt (1921) showed significant social class differences, particularly in the more verbal items of the Binet scales, while Gordon (1923), Wheeler (1942), Baldwin, *et al* (1930) and others have shown that the conventional intelligence test was unsuitable for children without formal education, from isolated communities or from rural districts in general.

Amongst the causal interpretations made of social class differences in measured intelligence major emphasis has usually been placed on hereditary and environmental influences. While not denying these, it is considered that the importance of test bias requires further study. Eells (1951) has produced strong evidence to support the contention that bias in intelligence test items is common, and that it tends to favour upper-middle class children. From his Chicago surveys, he concluded that verbal tests and items show significantly more variation in discriminating between social classes than do non-verbal.

Havighurst, *et al* (1944, 1947) also demonstrated that the extent of differentiation amongst social classes depends on the trait measured, the greatest discrepancies appearing in verbal and numerical tests, the smallest in spatial, mechanical and other non-verbal tests.

These research studies are, unfortunately, open to the criticism made by Lorge (1952) that analysis in terms of social levels reveals nothing about the ability of tests to differentiate between intellectual levels. Both dimensions are important in test evaluation. Moreover, Eells used only one non-verbal test whose reliability was suspect (Tyler, 1953), while Havighurst's studies were based on relatively small samples.

'Culture-fair' tests such as those of Cattell (1940, 1958) and the Davis-Eells Games (1953) have not adequately fulfilled their designers' intentions according to studies described by Anastasi and Cordova (1953), Drake (1955), Smith (1956) Stilwell (1958) and Love and Beach (1957), although the evidence has sometimes been more favourable (Marquant and Bailey 1955, Cattell 1958, Fowler 1955). Further evaluation of such tests using a construct validity approach in a variety of settings seems necessary in order to assess the usefulness of such measures.

Since the present investigators are seeking to identify tests suitable for a variety of cultural and educational backgrounds, it was considered that Cattell's approach to the problem of reducing cultural differentials by holding them to a minimum would be more suitable for this study than the Davis-Eells approach of holding them constant. On such an assumption, this investigation was designed to study tests similar to Cattell's in an attempt to identify test items which tap fundamental processes of intellectual ability while holding environmental differences to a minimum.

No illusions are held about the possibility of eliminating cultural influences from a pencil-and-paper test. Every culture stimulates the development of certain abilities at the expense of others, and the idea of an intelligence test being culture-free in the sense that it measures the products of heredity quite apart from the impact of cultural experience cannot be taken seriously. The present study aims rather to identify and evaluate tests which reduce the dependence on acquired knowledge in order to make possible the improvement of mental testing for children whose backgrounds do not fall within the required range assumed by the conventional verbal test.

III.—THE CONCEPT OF INTELLIGENCE.

Over the past several decades the concept of intelligence as general innate ability has been subjected to a barrage of criticisms. Factor analysts such as Thurstone (1938) and Guilford (1956) have disparaged the idea of a general factor; environmentalists such as Halsey (1959) and the Iowa investigators Stoddard and Wellman, 1940) have questioned the notion that intelligence is predetermined or fixed; developmental research pioneered by Piaget (1947) suggests that intelligence is intimately bound up with the continuous interaction between organism and environment, while the possibility of critical periods and irreversible influences in mental development is now seriously held (Hebb, 1958).

One beneficial effect of these criticisms has been to draw attention to the distinctions drawn by Hebb (1949) and Vernon (1955) amongst Intelligences A, B and C. Nevertheless, behavioral definitions of intelligence are as controversial as ever, and the operational approach of defining intelligence in statistical terms appears the more profitable course of action. Such a position was adopted in forming the conceptual framework of this study.

The hierarchical conception of intelligence described by Burt (1949) and Vernon (1950) is regarded by the present authors as the most useful model for assessing intelligence in diverse cultural settings. The general factor is a crucial construct for assessing the extent to which culture-reduced tests are sampling the common denominator of abilities required for success in a variety of mental tests. It looms so large in psychometric surveys of heterogeneous children that it is difficult to ignore, while the advantage of distributing its variance over a number of lower-level factors has yet to be demonstrated even with culturally homogeneous children. In predicting to distant and general criteria, when adaptive treatments are possible, 'pure' factor tests, are seldom superior to general intelligence tests, while the amount of time and money consumed on the former is an obvious administrative disadvantage.

The question of the innateness of the general factor can be circumvented by insisting only on its comparative stability in relation to lower-order factors. Evidence to support such stability comes from various sources. Woodrow (1938) demonstrated that the general factor was relatively unaffected by differential practice in a thirty-three variable analysis, while group factors changed considerably. Hofstaetter's (1954) factor analysis of the longitudinal mental development of children in the Berkeley Growth Study revealed a stable general factor increasing in size between the ages of 4 and 8, and remaining steady throughout pre-adolescence. Tests with high g loadings and negligible group factors have shown less susceptibility to cultural change (Cattell, 1958) and other environmental influences (Nisbet 1953, Burnett 1958).

Theoretically, the stable g factor is compatible with the hypotheses of Hebb (1949) and Ferguson (1956) which suggest that abilities are developed primarily by positive transfer from the old learning to the new. Early superiority in processing and integrating skills should lead to increasingly broad and cumulative superiority of intellectual development in a variety of fields. Vernon (1960), too, sees support for the general construct in the developmental approach for " those who, early in life acquire a larger stock of perceptual schemata and verbal habits are better able to build up more complex and more flexible schemata necessary for conceptual thinking " (p. 41). So the general all-inclusive logical structures accumulate and proliferate, producing a broad general factor which, although difficult to define behaviourally, continues to account for most of the statistical variance in any battery of ability tests.

In a comprehensive review of theories and research on intelligence, Hunt (1961) draws attention to the compatibility of apparently diverse interpretations of intelligence, which seem to be best explicable in hierarchical terms.

" A hierarchical factor structure, with Spearman's g at the apex, as the English factor analysts have tended to conceive of intelligence is precisely what would be expected from considering abilities as the accumulative effects of the transfer of training or learning in varied situations . . . Piaget's conception of accommodation corresponds to the change in function that occurs with training or learning, and Piaget's conception of assimilation corresponds to the notion of transfer of training " (Hunt, 1961, p. 359).

This general factor at the top of the hierarchical model, is essential to the present study. To qualify as a measure of an all-embracing intellectual ability, while relatively free from the vagaries of environmental circumstances, a test must show a substantial loading on the general factor at the top of the hierarchy. Culture reduction is of little value if the test does not sample extensively that common denominator or ' transfer basis ' of the fluid set of abilities we understand by intelligence.

IV.—EXPERIMENTAL PROCEDURE.

Nine tests and subtests were selected as promising culture-reduced measures of general intelligence. In May, 1960, these were given to 271 pupils constituting a stratified sample from the Grade VII population in the public schools of Edmonton, a rapidly expanding central Alberta community of nearly 300,000 people. The variables of provincial examination averages, socioeconomic status, geographic location, school type and residential mobility were considered in selecting the sample, and a comparison with a random selection from the same population in terms of age, sex, occupational status and Laycock I.Q., showed no significant differences.

The potentially useful culture-reduced tests selected were:

(1) Raven Standard Progressive Matrices (1956).
(2) Cattell Test of g "Culture Free." Scale 2. (1958).
(3) Lorge-Thorndike Non-Verbal Intelligence Test. Level 4 (1957).
(4) Lorge-Thorndike Figure Classification Subtest.
(5) Lorge-Thorndike Number Series Subtest.
(6) Lorge-Thorndike Figure Analogies Subtest.
(7) Holzinger-Crowder Uni-Factor Test (1952) Figure Changes.
(8) Holzinger-Crowder Uni-Factor Test (1952) Mixed Series.
(9) California Test of Mental Maturity (CTMM) Non-Language, Short Form, Elementary (1957).

Other variables for which data were gathered were:

(10) California Test of Mental Maturity, Language (1957).
(11) Holzinger-Crowder Uni-Factor Spatial Test (Boots) (1952).
(12) Occupational Index Score.
(13) Home Index Score.
(14) Year-end marks on teacher-made tests of attainment.

In May of the previous year, 1959, the same pupils had been administered the California Achievement Battery, Junior High (1957), whose subtests, listed below, were used to help identify factors in the factor analytic studies.

(15) California Reading Vocabulary.
(16) California Reading Comprehension.
(17) California Reading Total.
(18) California Arithmetic Reasoning.
(19) California Arithmetic Fundamentols.
(20) California Arithmetic Total.
(21) California Language.
(22) California Spelling.
(23) California English Total.

Further data used for comparative purposes were:

(24) Laycock Mental Ability Test (1935)—raw scores obtained two years earlier.
(25) Raven Coloured Progressive Matrices (1947)—administered in May, 1956.
(26) CTMM Primary (1953)—administered in May, 1956.

The CTMM Elementary (1957) was also analysed by subtests.

(27) CTMM Spatial.
(28) CTMM Logical.
(29) CTMM Number.
(30) CTMM Verbal.

The occupational index was obtained by assigning each pupil's statement (checked in a local directory) a standard score from Blishen's (1958) Canadian Socio-Economic Scale, based on Dominion census data, and evenly weighted according to income and years of training required. The Home Index Scale consisted of twenty items, adapted for local conditions from a scale devised by Gough (1949). The purpose of these measures was not to isolate the crucial variables operating to accelerate mental maturity, but rather to indicate when a number of these stimulating correlates was present and operating directly or indirectly, and when their importance was reduced.

On the basis of previous research, and the theoretical position adopted, the following predictions were made:

(*a*) The nine culture-reduced tests will show g loadings as high as those of the conventional verbal tests of intelligence.

(*b*) The nine culture-reduced tests will show negligible loadings on those group factors which appear to depend heavily on cultural acquisitions.
(*c*) Socioeconomic status of father will show greater correlations with scores on the conventional verbal tests than it does with scores on the culture-reduced tests.
(*d*) Over a period of four years (1956 to 1960) children from upper socioeconomic levels will improve their status on the CTMM more than on the culture-reduced Progressive Matrices.
(*e*) An item analysis of the most promising culture-reduced tests will show that upper and lower thirds, by socioeconomic status, will succeed in similar ratios on each item. On the CTMM Verbal Test, discrimination indices will show greater variation from item to item.

Additional evaluative criteria of relationship with school achievement, reliability and suitability for classroom use were also examined.

V.—ANALYSIS OF RESULTS.

Three socioeconomic variables were used in the analyses—occupational T-score, home index raw score, and a combination of the two measured with equal weights, yielding what was termed a socioeconomic index. Preliminary analysis on the two component sub-scores showed an intercorrelation of 0·61 which is close to Gough's corresponding coefficient of 0·65 (1949). The pattern of correlations with other variables was almost identical for the two measures (rho=0·87) suggesting that they sample highly similar correlates of mental ability. A split-half reliability coefficient of 0·77 was obtained from the Home Index questionnaire.

(*a*) and (*b*) The two hypotheses concerning the abilities sampled by each test were investigated by performing two factor analyses using Hotelling's Principal Components Method. The first analysis was performed on twenty-one variables (Tables 1 and 2) and included many subtests. The second concentrated on the eleven total test scores (Tables 3 and 4).

From the first matrix six principal components were extracted* but the last was discarded because it failed to reach significance level, according to the tests of McNemar, Burt, and Guilford and Lacey. The five significant factors were then rotated using criteria of positive manifold, maximization of the general factor and psychological meaningfulness. After four rotations a compelling psychological factor pattern emerged (Table 2). The factors were interpreted as general ability (g), a verbal educational factor (v : ed), socioeconomic status (se), numerical ability (n) and, possibly, speed (sp). The fact that no spatial factor appeared in the analysis may be due to the low ceiling of the CTMM Spatial Subtest, or the strong speed element in the Holzinger-Crowder Test. The eleven total test scores were analysed similarly yielding three clear factors labelled g, v : ed and n. A comparison of the factor patterns of the tests common to both analyses showed a mean loading discrepancy of only ·06, which suggests a high degree of similarity in the common dimensions obtained. The rotated factor solutions indicate that the tests most highly loaded with the general factor were Cattell's Test of g, the Progressive Matrices, and the Lorge-Thorndike Figure Analogies Subtest. The remaining six culture-reduced tests showed less promising g saturations, or significant loadings on verbal or numerical factors.

---

* Communalities were estimated by calculating the squared multiple correlation of each variable with the n-l remaining variables in the battery, as suggested by Wrigley (1959) and Kaiser (1960). Iteration was performed after replacing the communality with a value increased (or decreased) by one-and-a-half times the discrepancy between obtained and estimated values, until convergence occurred.

TABLE 1

CORRELATION MATRIX FOR TWENTY-ONE VARIABLES.
(Decimal Points Omitted.)

Test No.	1	2	4	5	6	7	8	11	12	13	15	16	18	19	21	22	24	27	28	29	30
1	(542)*																				
2	637	(629)																			
4	439	441	(358)																		
5	457	422	387	(494)																	
6	566	555	485	403	(556)																
7	494	457	347	388	379	(442)															
8	358	387	338	532	396	451	(555)														
11	417	393	237	247	262	356	277	(328)													
12	165	189	113	181	240	143	239	081	(552)												
13	237	244	152	176	233	237	302	168	610	(582)											
15	301	254	225	334	333	334	427	176	265	328	(742)										
16	466	387	280	405	405	427	461	161	250	303	681	(645)									
18	349	360	308	368	365	316	377	215	177	234	467	465	(371)								
19	396	303	335	435	362	358	470	130	118	162	447	586	459	(621)							
21	404	341	296	363	340	405	500	224	179	268	656	653	427	532	(647)						
22	172	068	105	143	207	114	193	-037	132	154	492	438	330	425	450	(445)					
24	507	498	435	438	526	428	451	268	309	324	643	662	489	479	621	441	(740)				
27	502	465	373	369	471	284	342	310	059	060	145	213	287	233	250	026	308	(437)			
28	527	493	362	383	455	372	373	284	286	239	355	459	313	317	372	184	550	382	(483)		
29	538	466	318	510	448	425	471	228	276	282	364	405	418	510	402	276	554	349	517	(571)	
30	337	352	291	340	374	353	425	136	286	383	717	503	438	421	580	449	667	187	469	429	(700)

* Final estimate of communality.

TABLE 2

ROTATED FACTOR SOLUTION AND COMMUNALITIES FOR TWENTY-ONE VARIABLES.
(Loadings below ·20 omitted)

Test No.	Test Name	I_3	II_2	III_1	IV_1	V_1	L^2
1	Prog. Matrices	708	—	—	—	—	542
2	Cattell g	754	—	—	—	222	626
3	L-Th. Fig. Class	582	—	—	—	—	355
5	L-Th. No. Series	547	—	—	395	—	486
6	L-Th. Fig. Anal.	738	—	—	—	—	553
7	Holz.-Crowd. Fig. Ch. ..	498	—	—	211	369	440
8	Holz.-Crowd. Series	457	214	—	453	303	554
11	Holz.-Crowd Spatial	397	—	—	—	396	326
12	Occ. Status	251	—	683	—	—	583
13	Home Index	247	205	669	—	—	602
15	Reading Vocab.	335	743	—	—	264	747
16	Reading Comp.	501	618	—	—	—	681
18	Arith. Reasoning	460	341	—	—	—	369
19	Arith. Fund.	450	436	—	442	—	629
21	Language	417	589	—	—	299	641
22	Spelling	201	618	—	—	—	443
24	Laycock	676	509	—	—	—	735
27	CTMM Spatial	612	—	—	—	—	428
28	CTMM Logical	660	—	—	—	—	476
29	CTMM Number	639	201	—	331	—	563
30	CTMM Verbal	458	658	—	—	—	694
% of Common Variance		51·18	25·23	9·11	7·80	6·68	100·00
% of Total Variance		27·94	13·78	4·97	4·26	3·65	54·60

TABLE 3

CORRELATION MATRIX FOR ELEVEN TOTAL TEST SCORES.
(Decimal Points Omittted.)

Test No.	1	2	3	7	8	9	10	17	20	23	24
1	(620)*										
2	637	(645)									
3	629	614	(666)								
7	494	457	470	(555)							
8	358	387	531	451	(555)						
9	577	505	605	384	460	(515)					
10	473	473	549	434	492	443	(727)				
17	425	356	467	419	486	351	772	(800)			
20	435	369	532	393	504	425	562	629	(574)		
23	381	297	408	363	466	353	619	733	598	(660)	
24	507	498	601	428	451	431	732	713	550	642	(741)

* Final estimate of the communality.

TABLE 4

ROTATED FACTOR SOLUTION AND COMMUNALITIES FOR ELEVEN TOTAL TEST SCORES.
(Loadings below ·20 omitted).

Test No.	Test Name	I_2	II_1	III_1	L^2
1	Prog. Matrices	780	—	—	634
2	Cattell g	792	—	—	639
3	Lorge-Thorndike	740	—	325	678
7	Holz.-Crowd. Fig. Ch.	521	—	256	374
8	Holz.-Crowd. Series	399	338	531	556
9	CTMM Non-Lang.	624	—	349	516
10	CTMM Language	580	616	—	720
17	Calif. Reading	444	784	—	819
20	Calif. Arith.	430	509	343	562
23	Calif. Eng.	361	702	—	654
24	Laycock	634	573	—	730
	% of Common Variance	55·97	32·83	11·20	100·00
	% of Total Variance	35·02	20·55	7·01	62·57

(*c*) To investigate the third hypothesis, the combined index of socioeconomic status was correlated with each test, and the difference between coefficients tested for significance. Table 5 shows the correlations of socioeconomic status with the nine culture-reduced tests in the left column and four verbal tests across the top. Significance levels of difference between coefficients are found in the body of the table.

Only three of the observed differences failed to reach the ·05 significant level, a finding which supports the hypothesis that the culture-reduced tests selected show less differentiation between socioeconomic levels than do the conventional verbal tests. The three tests found to have highest g loadings showed satisfactory linearity of regression and so can be considered to meet the third criterion of culture-reduction quite adequately.

TABLE 5

SIGNIFICANCE LEVELS OF DIFFERENCES IN CORRELATION COEFFICIENTS BETWEEN SOCIO-ECONOMIC STATUS AND SELECTED TESTS.

	Test	CTMM Total	Laycock Total	CTMM Lang.	CTMM Verbal
Test	r with S.E.S.	·384	·354	·414	·382
Progressive Matrices....	·228	·01	·02	·01	·01
Cattell Test of g	·244	·01	·05	·01	·02
Lorge-Th. Fig. Class ..	·152	·01	·01	·01	·01
Lorge-Th. No. Series ..	·192	·01	·01	·01	·01
Lorge-Th. Fig. Anal. ..	·262	·01	·05	·01	·05
Lorge-Th. Total........	·269	·01	·05	·01	·05
CTMM Non-Lang.	·184	·01	·01	·01	·01
Holz.-Crowd. Series ..	·310	N.S.	N.S.	·05	N.S.
Holz.-Crowd. Fig. Ch. ..	·220	·01	·02	·01	·01

(*d*) The Progressive Matrices and CTMM tests were used to test the fourth hypothesis of change in relationship with socioeconomic status over four years. Since appropriate forms of the Matrices and CTMM had been administered to the sample in both Grades III and VII, it was possible to predict an increase in correlation with socioeconomic status for the more culture-loaded but not the culture-reduced test. The fact that the correlations between grade levels were similar for each test (·545 and ·471, respectively) made the assumption of similar degree of overlap in the functions measured on both occasions appear reasonable, while the same coefficients were considered low enough to allow those socioeconomic variables associated with test scores to operate.

Correlation coefficients with socioeconomic status were calculated for normalized T-scores on all four tests, and differences tested for significance.* Results showed an insignificant increase of 0·015 in the case of the Matrices, whereas the corresponding CTMM Total correlation increased from 0·215 in Grade III to 0·323, a difference which is significant at the ·05 level. As predicted, the abilities sampled by the Matrices test appear less susceptible to the cumulative effects of those socioeconomic variables which influence the mental development of children between Grades III and VII.

(*e*) To investigate the fifth hypothesis, that a culture-reduced test would show less variation than a culture-loaded test in terms of item discrimination indices across socio-economic levels, the Progressive Matrices and CTMM Verbal Test were selected as representative of each category. The percentage passing each item in high and low socio-economic thirds was calculated for each test, distributions of discrimination indices drawn up for each test, and their variances compared. Only items with difficulty indices between 0·10 and 0·90 were included in the analysis.

The variances of the two distributions for the Matrices and CTMM Verbal Tests were calculated as 16·32 and 35·28, respectively, which, when compared, yielded a significant F ratio of 2·16 with 28 and 35 d.f. As hypothesized, items in the verbal test show greater variation in their discrimination between some socioeconomic levels than those of the non-verbal test. Although these discrepancies may not mar the predictive validity of the test assuming fixed treatments, they suggest that the verbal test is less suitable as a measure of intellectual potential. The Matrices discrimination indices were found to be similar throughout a wide range of difficulty.

The criterion of relationship with school marks and standardized tests was then investigated. End-of-year examination averages over all school subjects for each of the thirteen classes in the study were correlated with each of the intelligence tests and subtests. Results from the California Achievement tests administered in the previous year were also correlated with the major variables, using the whole sample of 271 pupils. The median correlations for the thirteen school marks and the mean correlations for the three California Achievement subtests are included in Table 6. Although less efficient than the culturally weighted conventional tests in this ‘ cold-blooded ’ predictive situation, the correlations of the better culture-reduced tests are substantial enough to indicate promise for adaptive treatments.

Typical reliability coefficients reported from previous research are included in Table 6. A check was made on the Progressive Matrices in view of its promise in terms of the other criteria. A split-half reliability co-efficient of 0·91 and a Kuder-Richardson coefficient of 0·83 indicated a degree of reliability very close to the average value reported in other studies at this age level. (Burke, 1958).

---

* The sample was reduced from 271 to 243 to test this hypothesis due to incomplete test data from Grade III.

All tests were considered convenient for classroom use, with the possible exception of the Holzinger-Crowder subtests, whose administration instructions were found to be tedious and repetitive, and time limits somewhat unrealistic. The Matrices test was judged the most popular by pupils and its ease of administration could certainly be regarded as an asset.

VI.—DISCUSSION AND CONCLUSIONS.

The main findings of the study have been drawn together in Table 6. In interpreting this table it should be remembered that, assuming adaptive rather than fixed treatment, the criteria of high g loading and low correlation with socioeconomic status should be accorded more weight than the present relationship with school achievement.

Table 6 indicates that the Progressive Matrices best meets the criteria chosen for evaluation of a culture-reduced intelligence test. In addition to the high g loading and low correlation with socio-economic status, this test is a reasonably good predictor of school marks, has a respectable reliability, shows no change in relationship with socioeconomic status over four years, and gives little indication of variation in item discrimination across socioeconomic levels. Moreover, the test has face validity in that the progressive increase in difficulty makes it a tester of what it teaches and it appears to be little affected by the previous experience of the testee.

The shorter Cattell Test of g and Lorge-Thorndike Figure Analogies Subtest rank next. Reorganization of instructions, lengthening, and item analyses might well improve these tests, particularly in relation to reliability and correlation with schoool marks, where they are weakest. The remaining culture-reduced tests showed less promise in this investigation although further study is currently being conducted on their potential values in relation to racial and language variables.

TABLE 6

SUMMARY OF SELECTED INTELLIGENCE TESTS EVALUATED AGAINST SIX CRITERIA.

Test	g loading	v:ed or n	r with S.E.S.	Median r Sch.Marks	Mean r Acht. Tests	Split-Half Reliability
Prog. Matrices	·78	—	·23	·56	·41	·91
Cattell g..........	·79	—	·24	·34	·35	·85*
Lorge-Th. Fig. Class	·58	—	·15	·31	·31	·71
Lorge-Th. Series ..	·55	·40(n)	·19	·45	·41	·84
Lorge-Th. Fig. Anal	·74	—	·26	·42	·39	·87
Lorge-Th. Total ..	·75	·33(n)	·27	·43	·47	·93
Holz.-Crowd. Series	·46	·45(n)	·31	·51	·49	—
Holz.-Crowd. Fig.Ch	·52	·21(n)	·22	·40	·39	—
CTMM Non-Lang.	·62	·35(n)	·18	·38	·38	·91
CTMM Lang.	·58	·62 (v:ed)	·41	·67	·66	·95
CTMM Total	—	—	·38	·66	·65	·95
Laycock Total ..	·68	·57 (v:ed)	·35	·59	·64	·95

* Re-test reliability coefficient.

For further research, too, one might predict considerable potential for an unspeeded test which used the Matrices approach of testing what it taught, and which supplemented the abilities measured with the most discriminating figure analogy and perceptual series items, or variations thereof.

In summary, it seems reasonable to conclude that as tests of intellectual potential, the conventional verbal intelligence tests examined in this study show discrimination against children from lower socioeconomic backgrounds. It has been shown that, for 12 and 13-year-old pupils :

(1) Culture reduced tests sample the general intellectual ability factor as well as or better than conventional tests.
(2) Most culture-reduced tests show negligible loadings on verbal and numerical factors.
(3) Culture-reduced tests show significantly less relationship with socio-economic status than do conventional tests.
(4) A conventional verbal test (CTMM) showed significant increase in relationship with socio-economic status over four years, whereas the Progressive Matrices showed no change.
(5) Verbal items from the CTMM showed greater variation in item discrimination between social classes than did items from the Progressive Matrices.

Generalizing further, the three most promising tests identified by the study could profitably be put to greater use in the estimation of intellectual ability for adaptive teaching purposes, and particularly in talent identification and classification of pupils from underprivileged and foreign language bckgrounds, for distinguishing between mental defect and reading difficulties due to other causes, for holding intelligence constant in research experiments where verbal ability is irrelevant, and generally, as a supplement to verbal intelligence measures in the routine testing of children from all socioeconomic levels.

VII.—REFERENCES.

ANASTASI, A., and CORDOVA, F. A. (1953). Some effects of bilingualism upon the intelligence test performance of Puerto Rican children in New York City. *J. Educ. Psychol.*, 44, 1-18.

BALDWIN, B. T., FILLMORE, E. A., and HADLEY, L. (1930). *Farm Children.* New York : Appleton-Century-Crofts.

BINET, A., and SIMON, T. (1916). *The Development of Intelligence in Children.* (Translated by Kite, E. S.) Baltimore : Williams and Williams.

BLISHEN, B. R. (1958). The construction and use of an occupational class scale. *Canad. J. Econ. Polit. Sci.*, 24, 519-531.

BURKE, H. R. (1958). Raven's progressive matrices : Review and critical evaluation. *J. Genet. Psychol.*, 93, 199-228.

BURNETT, A. (1955). Assessment of intelligence in a restricted environment. Unpublished Ph.D. Thesis. McGill University.

BURT, C. (1922). *Mental and Scholastic Tests.* London : King and Son.

BURT, C. (1949). The structure of the mind : a review of the results of factor analysis. *Brit. J. Educ. Pyschol.*, 19, 110-111, 174-179.

CATTELL, R. B. (1940). A culture-free intelligence test. *J. Educ. Psychol.*, 31, 161-180.

CATTELL, R. B. (1958). *Handbook for the Culture-Free Intelligence Test, Scale* 2. Champaign, Illinois : IPAT.

CRONBACH, L. J., and GLESER, G. C. (1957). *Psychological Tests and Personnel Decisions.* Urbana, Illinois : University of Illinois Press.

CRONBACH, L. J., and MEEHL, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull*, 52, 281-302.

DAVIS, A., and EELLS, K. (1953). *David-Eells Test of General Intelligence or Problem Solving Ability.* New York : World Book Co.

DRAKE, R. M. (1953). Review of Cattell Culture-Free Test, in Buros, O.K. (Editor). *Fourth Mental Measurements Yearbook.* New Jersey : Gryphon Press.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., and Tyler, R. S. (1951). *Intelligence and Cultural Differences.* Chicago: University of Chicago Press.

Elley, W. B. (1961). A comparative analysis of the socioeconomic bias in selected intelligence tests. Unpublished Ph.D. Thesis, University of Alberta.

Ferguson, G. A. (1956). On transfer and the abilities of man. *Canad. J. Psychol.*, 10, 121-131.

Fowler, W. L. (1955). A comparative analysis of pupil performance on conventional and culture-controlled tests. Ph.D. Thesis, Ann Arbor, Michigan.

Gordon, H. (1923). Mental and scholastic tests among retarded children. London: *Bd. Educ., Educ. Pamphlet, No. 44.*

Gough, G. (1949). A short social status inventory. *J. Educ. Psychol.*, 40, 52-56.

Guilford, J. P. (1959). *Personality.* New York: McGraw-Hill.

Halsey, A. H. (1959). Class differences in general intelligence: I. *Brit. J. Stat. Psychol.*, 12, 1-4.

Havighurst, R. J., and Breese, F. H. (1947). Relation between ability and social status in a mid-western community: III. *J. Educ. Psychol.*, 38, 241-247.

Havighurst, R. J., and Janke, L. L. (1944). Relation between ability and social status in a mid-western community: I. *J. Educ. Psychol.*, 35, 357-368.

Hebb, D. O. (1949). *Organization of Behavior.* New York: John Wiley and Sons.

Hebb, D. O. (1958). *A Textbook of Psychology.* Philadelphia: W. B. Saunders Co.

Hofstaetter, P. R. (1954). The changing composition of 'intelligence': a study in T technique. *J. Genet. Psychol.*, 85, 159-164.

Hunt, J. McV. (1961). *Intelligence and Experience.* New York: Ronald Press.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educ. and Psychol. Meas.*, 20, 141-151.

Lorge, I. (1952). 'Intelligence and cultural differences' as seen by the test-makers. *Teach. Coll. Rec.* 54, 190-193.

Love, M. I., and Beach, S. (1957). Performance of children on the Davis-Eells Games and other measures of ability. *J. Consult. Psychol.*, 21, 29-32.

MacArthur, R. S. (1962). The construct validity of a number of measures of intellectual potential for an Alberta Metis sample. Paper presented at Annual Meeting of Canadian Psychological Association.

Marquart, D. I., and Bailey, L. L. (1955). An evaluation of the Culture-Free Intelligence Test. *J. Genet. Psychol.*, 86, 353-358.

Nisbet, J. D. (1953). Family environment: a direct effect of family size on intelligence. Occasional paper on Eugenics. London: Cassell and Co.

Piaget, J. (1947). *Psychology of Intelligence.* Translated by Piercy, M., and Berlyne, D.E. London: Routledge and Kegan Paul.

Smith, T. W. (1956). Comparison of test bias in the Davis-Eells Games and the CTMM. *Calif. J. Educ. Res.*, 7, 159-163.

Stern, W. (1914). *The Psychological Methods of Testing Intelligence.* Baltimore: Warwick and York.

Stillwell, L. (1958). A study of the Davis-Eells Games. Unpublished Master's Thesis. Ohio: Kent State University.

Stoddard, G. D., and Wellman, B. L. (1940). Environment and the I.Q. *N.S.S.E. Yearbook*, 39, (1), 405-442.

Thurstone, L. L. (1938). Primary Mental Abilities. *Psych. Monog. No.* 1. Chicago: University of Chicago Press.

Tyler, F. T. (1953). Comments on the correlational analysis reported in "Intelligence and Cultural Differences." *J. Educ. Psychol.*, 44, 288-295.

Vernon, P. E. (1950). *The Structure of Human Abilities.* London: Methuen and Co.

Vernon, P. E. (1955). 'The Assessment of Children' in The Bearings of Recent Advances in Psychology on Educational Problems. London: University of London Institute of Education.

Vernon, P. E. (1960). *Intelligence and Attainment Tests.* London: University of London Press.

Wheeler, L. R. (1942). A comparative study of the intelligence of East Tennessee Mountain children. *J. Educ. Psychol.*, 33, 321-334.

Woodrow, H. (1938). The relation between abilities and improvement with practice. *J. Educ. Psychol.*, 29, 215-230.

Wrigley, C. (1958). Objectivity in factor analysis. *Educ. and Psychol. Meas.*, 18, 463-476.

(*Manuscript received, October,* 1962.)

THE RELATION OF PSYCHOLOGICAL TEST SCORES TO ACADEMIC BIAS

By LIAM HUDSON

(*The Psychological Laboratory, University of Cambridge*)

Summary. On entering the Sixth Form, English schoolboys choose between Arts and Science subjects. Effectively, this choice commits them to one career field or the other. The present research is aimed to discover whether psychological tests reflect the academic bias of such boys towards Arts or Science subjects ; and whether they give any indication of boys' aptitude for specialised work at university and in research. A battery of conventional psychological tests was given to a sample of 315 15 and 16-year-old schoolboys, drawn from five schools. These boys had just finished the G.C.E. 'O' level examination, and stood a good chance of reaching university.

Results show that the tests used do reflect academic bias surprisingly well. On the other hand, there are signs that the relation of tests scores to aptitude for university and research work will prove to be complex. Some of the most gifted young Science specialists, for example, have scores which bear no relation to the trend of results in the sample as a whole. The implications of these and other findings are discussed.

I.—Introduction.

On entering the Sixth Form, at the age of 15 or 16, the clever English schoolboy has to choose between Arts and Science subjects. The choice he makes determines to a considerable extent the eventual course of his career. If he becomes, say, a History specialist, he will find it difficult later on to change to Classics or French ; and even more difficult, if not impossible, to change to Mathematics or Biology. Obviously, the more relevant information he has before him when making this decision, the better.

It is this choice which forms the context for the present research. Its ultimate aim is a series of tests which would show a 16-year-old (*a*) roughly what his own potentialities are, and (*b*) what demands would be made on him by various careers at university and in research. In order to develop and validate such tests, a series of longitudinal experiments are needed, in which individuals' test scores as schoolboys are related to their achievements as adults. The present paper is merely an interim report on the first stage of one such project. It concerns the basic question : do test scores reflect a schoolboy's bias of ability towards Arts or Science at the age when the crucial choice is being made ? Although this relation of test scores to academic bias will be discussed in detail, no attempt will be made to show that the relation is stable through time, or more generally that the tests have predictive validity. The criteria used are those appropriate to the Lower Sixth Form, and these are not grounds for vocational advice.

Relatively little is known about the relation of test scores to academic bias. During the last twenty years, many attempts have been made to validate differential aptitude batteries, but these have met with little success (Super, 1957 ; Horst, 1959 ; Anastasi, 1961). A notable exception is Crawford and Burnham's (1946) report that the Yale Differential Aptitude Battery has differential validity in predicting college grades. They also report (p. 242) that certain undergraduates have marked and persistent verbal/spatial biases of test score. Similar biases have been reported by Bradford (1948) among British 11-plus candidates. The results in both cases imply that such biases of test score are related to biases of academic ability and interest; but these relations have not been substantiated (Burt, 1950).

II.—METHOD.

(1) *The Sample.*

The present results are based on a sample of 315 15 and 16-year-old schoolboys. These were drawn from five schools: two public boarding, one public day, and two grammar. All five have high academic reputations. In each case, the boys were tested as they entered the Sixth Form, having just taken the G.C.E. 'O' level examination. They were selected as having a reasonable chance of reaching university. The general level of their ability was high, the mean intelligence test score being only a few points lower than that found among Cambridge undergraduates (Hudson, 1960).

(2) *Subject-Groups and Criterion-Groups.*

In analysing the results, two systems of classification are used: Subject-Groups and Criterion-Groups. In the first of these, boys are grouped in terms of their special subject: Classics, Modern Languages, History, Physical Science, or Biology. This method is straightforward and yields a number of interesting comparisons. On the other hand, it gives little indication either of a boy's aptitude for his special subject, or of his academic bias. The most useful analysis is one which relates test scores to criteria *both* of academic success *and* of academic bias. In the present case, this is done by using Criterion-Groups.

Both the academic bias and the academic success of each boy were assessed from questionnaires, staff reports and G.C.E. 'O' level results. Four Criterion-Groups were then selected, with twenty-five boys in each: 'Arts Specialists' consisting of boys who are highly successful in Arts subjects and poor at Science; 'Arts All-rounders" who are highly successful at Arts and adequate at Science; 'Science-All-rounders' who are highly successful at Science and adequate at Arts; and 'Science Specialists' who are highly successful at Science and poor at Arts. As they were restricted to boys who are highly successful in their own subject, these Criterion-Groups reflect both differences in academic bias and differences in academic level. Differences in academic *bias* are reflected in comparisons between the Criterion-Groups; differences in academic *success* are reflected in comparisons between the Criterion-Groups and the remaining 215 boys in the sample. Four Criterion-Groups rather than five were used because genuine all-rounders proved relatively hard to find. The number of boys in each of the Subject- and Criterion-Groups is shown in Table 1. (It is important to note that the two systems of classification are independent, in the sense that a Classicist may be found in either of the Arts Criterion-Groups or in Rest-of-Arts, but not, of course, among the Scientists.)

(3) *The Indices Used.*

General Intelligence. The test used was A.H.5, a group test of high-grade intelligence specially suited to student samples (Heim, 1956). The test is administered in two Parts: Part I being verbal and numerical, Part II being diagrammatic.

Accuracy. A number of accuracy scores were derived from A.H.5, the most useful being ' Worse Accuracy ' and ' Range Accuracy.' A boy's accuracy on Parts I and II of A.H.5 were calculated separately—the number of correct answers as a percentage of the number attempted. The lower of the two percentages was then defined as his Worse Accuracy, and the difference between the two as his Range Accuracy. The accuracy score given in Tables 2-5 is Worse Accuracy.

Spatial Ability. An experimental spatial test was used as none of those already published was sufficiently difficult (Heim and Watts, 1957). It was aimed to tax students' ability at manipulating geometrical shapes in their mind's eye, but in practice was found to correlate substantially with General Intelligence (r=approximately ·5).

Vocabulary. An experimental, multiple-choice form was used, published tests being too easy (Heim and Watts, 1961).

General Knowledge. An experimental open-ended test was used, involving factual knowledge about 100 men, famous in current affairs, literature, painting, music, science.

Cognitive Bias. A differential score expressed the bias of each boy's ability in the cognitive tests. This contrasted his Verbal ability on the one hand with his Numerical and Spatial abilities on the other. (These scores were derived from A.H.5, A.H.5 and the Spatial Test, respectively.)

Bias of Interests. Interests were assessed by a system of cross-checks. Five sources of information were used, although all five were not available in every case ; questionnaire, form-ratings, general knowledge, school staffs, interviews. A differential score was used to measure each boy's bias towards ' cultural and political ' as opposed to ' practical and outdoor ' interests. A working definition of this second category of interest was reached with some difficulty. Certain activities contribute heavily to an Arts/Science discrimination : car and radio mechanics, model building, carpentry, photography, camping, rock climbing, sailing, bird-watching. While others, ostensibly similar, show a significant tendency to obscure it : cine-photography, model railways ; hunting, shooting, fishing, cycling, skiing, foreign travel. Similar difficulties were found in the ' cultural ' sphere ; interests in literature and painting contribute heavily to an Arts/Science discrimination, whereas an interest in music tends to obscure it. As the present aim is primarily discriminative, an operational definition of these categories was felt to be in order. So those interests already mentioned as obscuring the Arts/Science discrimination were left out of account.

With one exception, the seven indices mentioned above are reasonably independent of one another, the correlations between them ranging from zero to ·5. The exception concerns the scale of Cognitive Bias and the Spatial Test from which it is partially derived, r in this case being of the order of ·7. Scores on the first five of these indices are graded on a five-point scale, A-E. Grade A represents the top 10 per cent. of each school sample ; B, the next 20 per cent. ; C, the middle 40 per cent. ; D, the next 20 per cent. ; and E, the bottom 10 per cent. A fuller account of tests and methods is given in Hudson (1961).

III.—RESULTS.

(1) *Subject-Groups.*

The first aim in presenting the results is to give test ' profiles ' for each Subject-Group ; to show, for example, whether there are characteristic differences between Classicists and Historians. In the interests of clarity, Table 2 summarises these profiles qualitatively. The over-simplification which this involves is relatively unimportant, because the difference between any given pair of Subject-Groups is expressed quantitatively in Table 3. Table 3 indicates which of the differences shown in Table 2 are of real discriminative importance.

Table 2 shows that Classicists are typically of average ability or better, and that they are particularly strong in accuracy and vocabulary. Their interests are diverse and they have no marked cognitive bias. Modern Linguists are unremarkable, apart from a verbal cognitive bias and, very surprisingly, a *weakness* in vocabulary—an inexplicably large number of highly successful Linguists had vocabulary grades of D or E. Historians are clearly distinguished by low cognitive scores and a strong ' cultural and political ' bias of interests. Physical Scientists have high scores on cognitive tests, and a marked bias both of cognitive ability and of interest. Biologists are undistinguished except for a strong ' practical and outdoor ' bias of interests.

TABLE 1

THE SAMPLE TESTED, DIVIDED IN TERMS OF (i) SUBJECT-GROUPS, AND (ii) CRITERION-GROUPS.

	SUBJECT-GROUPS	n	CRITERION-GROUPS	n
ARTS	Classics	27	Specialists	25
	Modern Languages	39	All-Rounders	25
	History	68	Rest-of-Arts	84
		134		134
SCIENCE	Physical Science	142	Specialists	25
	Biology	39	All-Rounders	25
			Rest-of-Science	131
		181		181

TABLE 2

SUBJECT-GROUPS : A QUALITATIVE SUMMARY OF TEST SCORES.

	GENERAL INTELLIGENCE	ACCURACY	SPATIAL ABILITY	VOCABULARY	GENERAL KNOWLEDGE	COGNITIVE BIAS	INTERESTS	n
CLASSICS	Fair	V. Good	Good	V. Good	Fair	None	Diverse	27
MODERN LANGUAGES	Fair	Fair	Fair	Weak	Fair	Verbal	Diverse	39
HISTORY	Poor	Weak	Poor	Fair	Fair	Verbal	Cultural & Political	68
PHYSICAL SCIENCE	V. Good	Good	V. Good	Fair	Fair	Numerical & Spatial	Practical & Outdoor	142
BIOLOGY	Fair	Fair	Fair	Weak	Fair	None	Practical & Outdoor	39

TABLE 3

INDICES WHICH DISCRIMINATE BETWEEN ANY GIVEN PAIR OF SUBJECT-GROUPS.

	Modern Languages	History	Physical Science	Biology
Classics …………	Vocabulary	Spatial Accuracy	Interests	Interests Vocabulary
Modern Languages ..		Intelligence	Cognitive Bias	—
History …………			INTERESTS Cognitive Bias Spatial Intelligence	INTERESTS
Physical Science ….				—

STRONG DISCRIMINATION=0%—15% errors.
MODERATE DISCRIMINATION=16%—25% errors.
Weak Discrimination=26%—35% errors.

In describing the discriminations between the various Subject-Groups, three standards of discriminative efficiency are established : Strong, Moderate and Weak. A discrimination between two groups is defined as Strong if the proportion of errors involved is 15 per cent. or less, as Moderate if the proportion of errors lies between 16 per cent. and 25 per cent., and as Weak if it lies between 26 per cent. and 35 per cent. For example : if 15 per cent. of Historians lie above a given cut-off score on the vocabulary test while only 25 per cent. of Classicists lie below it, the proportion of errors involved in discriminating between these two groups is 20 per cent., and the discrimination is defined as Moderate. These three categories are distinguished typographically in both Tables 3 and 5 ; and in each cell of these Tables, the indices are arranged in descending order of discriminative efficiency. The cut-off score used in each case is the median for the whole schoolboy sample. All the discriminations shown in Tables 3 and 5 are statistically significant ($P=0{\cdot}05$; X^2 Test with Yates's Correction).

Although none of the discriminations between Subject-Groups is Strong, Table 3 shows that there are a number of significant differences between boys who specialise in the various academic subjects. The most obvious is that separating Historians from the Physical Scientists. For the rest, Classicists are distinguished from Modern Linguists and Biologists by their high vocabulary and from both groups of Scientists by the bias of their interests. Modern Linquists are distinguished from Physical Scientists by their cognitive bias.

(2) *Criterion-Groups.*

The Criterion-Groups consist of boys who are highly academically successful ; and they differ only in their academic bias. So contrasts between Criterion-Groups should reveal whether academic *bias* is reflected in test scores ; and the contrast of the Criterion-Groups with the rest of the sample should reveal whether academic *success* is so reflected. The profiles of the four Criterion-Groups and the remainder of the sample are summarised qualitatively in Table 4.

In general, the Arts Specialists have poor cognitive scores. They are also distinguished from Scientists by their interests, and from all other Criterion-Groups by their cognitive bias. The other three Criterion-Groups tend to have high scores, and the differences between them are relatively slight. These lie in interests, cognitive bias and, perhaps, spatial ability. The two 'Rest' groups score rather similarly. It seems that they occupy middle ground between the Arts Specialists at one extreme and the remaining three Criterion-Groups at the other.

The efficiency of each index in distinguishing between any given pair of Criterion-Groups is shown in Table 5. Again, only statistically significant differences are mentioned. The results may be summarised as follows :

(*a*) On the whole, discriminations between Criterion-Groups are much stronger than those between Subject-Groups.

(*b*) Among *successful* schoolboys, one distinction is of over-riding importance : that which separates Arts Specialists (i.e., those who are poor at Science) from the rest.

(*c*) Relative to this gulf, the dichotomy between Arts and Science is less important. Only his interests reveal the direction in which a successful all-rounder has specialised.

(*d*) Academic bias is reflected excellently among Arts boys, but poorly among Scientists. Or—to express the same relation in reverse—the tests reflect weakness in Science excellently, but only poorly reflect weakness in Arts.

(*e*) The relative similarity of Arts Specialists and Rest-of-Arts groups, and the dissimilarity of both from Arts-All-rounders emphasizes that tests do not measure the qualities underlying success in Arts, and also that the Arts All-rounder is a relatively unusual type.

(*f*) There is a surprising, although non-significant, tendency for Arts Specialists' cognitive scores to be *worse* than those of the Rest-of-Arts group.

(*g*) The indices : General Intelligence, Accuracy, and Spatial Ability all distinguish boys who are good at Science from those who are poor. Nevertheless, very high Accuracy scores are largely restricted to *Arts* All-rounders, while the Accuracy scores of successful Scientists tend to cluster in the B and C grades. Vocabulary measures academic aptitude in Arts and Science alike ; and, to a lesser extent, the same seems to be true of General Knowledge. Cognitive Bias gives a good indication of academic bias, particularly among Arts boys ; and Bias of Interests distinguishes Arts boys as a whole from Scientists.

Tables 4 and 5 do not give a clear idea of how efficiently the various academic groups could be distinguished in practice, if all seven indices were combined. As a rough check on the strength of these discriminations, the scores of all the Criterion-Group members were placed on cards and an attempt was made to identify each boy in the light of the evidence given in Tables 4 and 5. The results, shown in Table 6 are most satisfactory :

(*a*) Two-thirds of the Criterion-Group members are identified correctly, whereas only some 25 per cent. would have been indentified correctly had they been sorted at random. As many as 97 out of 100 are placed either correctly or in a neighbouring group.

TABLE 4

CRITERION-GROUPS: QUALITATIVE SUMMARY OF TEST SCORES.

	GENERAL INTELLIGENCE	ACCURACY	SPATIAL ABILITY	VOCABULARY	GENERAL KNOWLEDGE	COGNITIVE BIAS	BIAS OF INTERESTS	n
ARTS SPECIALISTS	Poor	Poor	Poor	Good	Fair	Verbal	Cultural & Political	25
ARTS ALL-ROUNDERS ..	V. Good	V. Good	Fair	V. Good	Good	None	Cultural & Political	25
SCIENCE ALL-ROUNDERS	V. Good	Good	Good	V. Good	V. Good	None	No Bias	25
SCIENCE SPECIALISTS ..	V. Good	Good	V. Good	Good	Good	Numerical & Spatial	Practical & Outdoor	25
REST-OF-ARTS	Weak	Weak	Weak	Fair	Weak	None	No Bias	82
REST-OF-SCIENCE	Fair	Fair	Fair	Weak	Fair	None	Practical & Outdoor	133

TABLE 5

INDICES WHICH DISCRIMINATE BETWEEN ANY GIVEN PAIR OF CRITERION-GROUPS.

	ARTS ALL-ROUNDERS	SCIENCE ALL-ROUNDERS	SCIENCE SPECIALISTS	REST-OF-ARTS	REST-OF-SCIENCE
ARTS SPECIALISTS ..	ACCURACY COGNITIVE BIAS INTELLIGENCE Spatial	SPATIAL ACCURACY INTELLIGENCE Cognitive Bias Interests	*SPATIAL* *COGNITIVE BIAS* INTERESTS INTELLIGENCE Accuracy	—	INTERESTS COGNITIVE BIAS Spatial Accuracy Vocabulary
ARTS ALL-ROUNDERS		INTERESTS	INTERESTS	INTELLIGENCE Accuracy Vocabulary Cognitive Bias	*INTERESTS* VOCABULARY Intelligence
SCIENCE ALL-ROUNDERS			Cognitive Bias	INTELLIGENCE GENERAL KNOWLEDGE Vocabulary Accuracy Spatial	VOCABULARY General Knowledge Intelligence
SCIENCE SPECIALISTS ..				SPATIAL COGNITIVE BIAS INTELLIGENCE Interests Accuracy	Intelligence Spatial
REST-OF-ARTS					Cognitive Bias Interests

STRONG DISCRIMINATION = 0%—15% errors.
MODERATE DISCRIMINATION = 16%—25% errors.
Weak Discrimination = 27%—35% errors.

TABLE 6

AN ATTEMPT TO IDENTIFY CRITERION-GROUP MEMBERS, USING ALL SEVEN INDICES TOGETHER.

	Identified as—			
	" Arts Spec."	"Arts All-R"	" Sc. All-R"	" Sc. Spec."
	n	n	n	n
ARTS SPECIALISTS	18	6	1	—
ARTS ALL-ROUNDERS	3	19	2	1
SCIENCE ALL-ROUNDERS	1	5	14	5
SCIENCE SPECIALISTS	—	—	9	16

(*b*) There is some confusion between neighbouring groups, particularly between the two groups of Scientists; but the discrimination between the other groups is good, and that between the two extremes, the Arts and Science Specialists, is almost complete. Not one Arts Specialist was identified as a Science Specialist, or *vice versa.*

(*c*) The discrimination between Arts and Sciences is also surprisingly strong; only four Arts boys were identified as Scientists, and only six Scientists were identified as Arts boys.

The samples used here are rather small, and a certain degree of chance error is inevitable. When the research is extended to other schools, particularly those with different academic traditions, changes must be expected. (It has been found, for instance, that the intellectual calibre of Biology specialists varies considerably from one school to another.) It seems unlikely, on the other hand, that the major differentiations will be materially altered.

(3) *Other Findings.*

(i) *Gifted Scientists.* The general level of test scores among successful Physical Science specialists is high; and, in practice, most gifted young scientists are easily identifiable from their scores. There are, nevertheless, a few exceptionally gifted Physical Science specialists whose scores are unaccountably low. Perhaps three cases are sufficient to illustrate this type: D., a mathematical 'prodigy'; P., a Chemistry specialist who has since gained an Open Scholarship at Cambridge; and R., the best of a remarkable group of young scientists at the public day school. Their test scores are respectively: General Intelligence—C, D, C; Accuracy—E, C, D; Spatial Ability—A, D, C; Vocabulary—B, B, C; General Knowledge—C, B, A. All three have strong or fairly strong interests, but none has either a 'political and cultural' or 'practical and outdoor' bias. All three are considered by their form-mates to be hard-working. In general, their remarkable talents reflect themselves in their test scores in hardly a single respect.

Although such cases account for a mere 2–3 per cent. of the whole sample, they demand an explanation. The most obvious is that the boys concerned were not trying. With most of them, this can be ruled out of account. Most of them are notably conscientious; and as their academic reputation usually preceded hem, they were watched with special care during the testing sessions. A second possibility is that these boys were made over-anxious by the novelty of the testing situation. Again, this seems feasible in a few cases but not in all. There are marked differences of personality and emotional maturity within this group, so it is unlikely that a single emotional factor could account for all of them. A third and much more likely explanation is that there are important scientific abilities which the present tests miss. There is, after all, no *prima facie* reason for believing that the complexity of scientific work can be epitomised in the present handful of paper-and-pencil puzzles. Indeed, it would be strange if it could. There are several normal scientific skills which conventional tests take little account of ; the ' inductive ' skills of marshalling complex evidence, for example, and accurate generalisation from incomplete evidence. Also there is the whole area of creativity (e.g., Getzels and Jackson, 1962). If new testing methods are needed, the best way of discovering them is presumably a detailed study of present anomalies.

(ii) *Extreme Arts Specialists.* Another tendency obscured in the general presentation of results concerns a small group of Arts Specialists. These are boys who are outstanding at Arts subjects but very weak at subjects which tax their powers of formal reasoning, e.g., Mathematics, Physical Science, Latin. This group shows a strong *negative* relation between such measures as General Intelligence, Accuracy, Spatial Ability on the one hand, and General Knowledge on the other. The scores of three of these extreme Arts Specialists—C., G., S.—are : General Intelligence—E, E, E ; Accuracy—E, E, E ; Spatial Ability—E, E, E ; Vocabulary—B, C, B ; General Knowledge—A, A, A. All three have a very strong ' political and cultural ' bias of interests, and a heavily verbal bias of cognitive ability. Both C. and G. are exceptionally hard-working. The General Knowledge scores of these three boys are, respectively, the fourth highest, third highest and the highest in the whole sample ; while, with one exception, G. has the lowest General Intelligence in the whole sample, and C. the lowest but one. As with the Physical Scientists mentioned above, this type is exceptional ; such boys account for less than 5 per cent. of the whole sample. Yet, they are possibly of even greater practical importance, because it is precisely this type of boy who is filtered out of the academic system by university requirements in Latin and Mathematics.

(iii) *Negative and Compensatory Relations.* This negative relation among extreme Arts Specialists seems to be one instance of a more general pattern. Among Historians and Modern Linguists there is a marked tendency for a high score on any one test to be counter-balanced by low scores on others. In cases where a boy's scores were all low, he was found almost invariably to be exceptionally hard-working. So it seems that in History and Modern Languages certain mental abilities are mutually inter-changeable ; and that where ability is lacking, hard work will serve in its stead. It will be interesting to see which of the various test profiles found among Arts boys bode well academically ; and whether qualitative differences of intellectual type are detectable between them. Why this form of equivalence should not hold among Classicists, Physical Scientists or Biologists is not clear. One possibility is that History and Modern Languages admit a wide range of differences in intellectual ' style,' whereas in Classics and Science there may be a greater tendency for one intellectual method to be the best.

These results suggest that a compensatory mechanism may be at work in some of the boys concerned ; that special weakness in, say, formal reasoning has caused them to concentrate on general knowledge. Compensation of this kind is illustrated in a case of sibling rivalry. These two brothers were 18 months apart in age, yet the younger had overtaken the older academically. They had an unsatisfactory home background; and at the time of his interview, the elder brother showed obvious signs of emotional disturbance. The younger brother was a Physical Scientist with cognitive scores above the average, a spatial bias of abilities, poor general knowledge and exclusively practical interests. The elder was a Historian with very low cognitive scores, a verbal bias of ability, good general knowledge, and exclusively cultural interests. In effect, the elder boy had concentrated his efforts in precisely those areas where his brother was weak.

(iv) *Accuracy, Emotional Stability and Originality.* Previous work has shown that accuracy scores can be used as an index of emotional stability (Petrie, 1948). In the present research, both Worse and Range Accuracy were found to correlate substantially with interview ratings of emotional stability (Hudson 1961)—r=·42 and ·65, respectively; n=51. Individual cases also suggested that there may be a connection between accuracy scores and potential originality. The Worse Accuracy scores of the six exceptionally gifted boys mentioned above were poor. The same is true of their Range Accuracy, their grades being E, C, D and E, E, D for Arts Specialists and Physical Scientists, respectively. Among Physical Scientists such poor accuracy scores are relatively rare, and generally are a sign of academic mediocrity. Although superficially anomalous, this finding is reminiscent of some previous suggestions about the nature of originality. Guildford (1950), for example, has suggested that creativity may stem from a general distaste for repetition; and Getzels and Jackson (1962) found 'creative' students impatient with the routine of classroom work.

IV.—CONCLUSION.

With certain important reservations, we may conclude that the academic bias of clever schoolboys does reflect itself in test scores with surprising clarity. If this relation holds true at university and in research, the present tests could be of some use in advising schoolboys about their choice of career. On the other hand, there are good reasons for expecting this relation to weaken. The demands made by a particular subject, Physics for example, may well change radically between G.C.E. 'O' level and Tripos; work for Tripos may differ from research; and the various branches of research in an area like Physics may differ among themselves (Roe, 1953). So although conventional tests may be said to have proved their worth at the Sixth Form level, the influence of ill-defined qualities like originality and maturity of judgment may well increase as the present sample matures; and if they increase, conventional tests may tend to lose their value.

Even at the Sixth form level, much remains to be done. The present tests do not reflect ability in Arts subjects, nor do they distinguish certain exceptionally gifted young scientists. New tests are certainly needed, but there is little indication of what shape they should take. The abilities which the present tests fail to detect may be primarily cognitive, e.g., the ability to marshall complex evidence. They may be more personal, e.g., imaginativeness, maturity of judgment, insight into human motives. Or they may involve differences of intellectual attitude and 'style' (Rapaport 1957; Getzels and Jackson 1962).

The fact remains that there are brilliant boys with relatively poor test scores, and seemingly dull boys whose scores are relatively high. As Getzels and Jackson point out, conventional psychological tests normally account for less than a quarter of the variance in academic success; and it seems that the deductive, puzzle-solving abilities which these tests do measure have little bearing on intellectual originality. So although there is scope for new cognitive tests of the orthodox kind, the most fruitful research may well lie in those areas of attitude and 'style' where cognitive and personal qualities interact. In the past there has been a tendency to approach these areas obliquely, using projective techniques like the Rorschach (Rapaport 1957.) ; but distinctions of type emerging from this work have not yet been related to differences of attitude and 'style' discernible in academic practice. Obviously, if these distinctions are to stand any chance of explaining intrinsically interesting phenomena, this gulf between the laboratory and the world at large will have to be bridged. It is at this point that the skills of the natural historian become vital. If test scores are to be related to the real world—and this, after all, is what validation should consist in—mental testers are obliged to observe intellectual workers at work, to analyse precisely what they do and how they do it. The immediate task, in other words, is to find ways of describing and quantifying those intellectual qualities which do not lend themselves to quantification automatically. A good example of the exploratory work which this will involve is Getzels and Jackson's (1962) study contrasting 'highly intelligent' and 'highly creative' school children. This study not only shows sensitivity to *qualitative* differences in school children's thinking, but also tackles the formidable problems of description which these qualitative differences present. Perhaps it is a good omen for the future of mental testing that the statistics these authors use are largely rudimentary.

Finally, the limitations of the present results must be reasserted. In the first place, the intellectual demands made by the various Sixth Form subjects are relatively elementary ; in the second place, the choice between them by no means determines the course of a schoolboy's career. A Sixth Form History specialist may end up as an accountant, businessman, politician, anthropologist, soldier, barrister, clergyman. He may have talents unrelated to his job—he may work in a bank and write poetry, or in a customs house and paint. Or he may have talents which circumstances force him to neglect. The present categories are bound to be provisional, and no really satisfactory assessment can be made until the present sample has been followed through university and beyond. Their test scores as schoolboys can then be related not merely to formal criteria like degree class, but to full descriptions of their intellectual lives as adults.

ACKNOWLEDGEMENTS.—I am deeply indebted to the headmasters and members of staff whose co-operation has made this work possible ; to Professor O. L. Zangwill, Professor P. E. Vernon and Dr. A. W. Heim for their help and advice ; to Dr. A. W. Heim for making available unpublished testing material ; and to the Ministry of Education and Department for Scientific and Industrial Research who have given the work their financial support.

V.—REFERENCES.

ANASTASI, A. (1961). *Psychological Testing*, 2nd edit. New York : Macmillan.

BRADFORD, E. J. G. (1948). Symposium on the selection of pupils for different types of secondary schools. *Brit. J. Educ. Psychol.*, **18**, 67.

BURT, C. (1950). Symposium on the selection of pupils for different types of secondary schools. *Brit. J. Educ. Psychol.*, **20**, 1.

CRAWFORD, A. B., and BURNHAM, P. S. (1946). *Forecasting College Achievement*. Yale : University Press.

GETZELS, J. W., and JACKSON, P. W. (1962). *Creativity and Intelligence*. New York : Wiley.

GUILDFORD, J. P. (1950). Creativity. *American Psychologist*, **5**, 444.

HEIM, A. W. (1956). *Manual for the Group Test of High Grade Intelligence A.H.5.* London: National Foundation for Educational Research.

HEIM, A. W., and WATTS, K. P. (1957). An experiment on practice, coaching and discussion of errors in mental testing. *Brit. J. Educ. Psychol.*, 27, 199.

HEIM, A. W., and WATTS, K. P. (1961). A preliminary study of the self-judging vocabulary scale. *Brit. J. Psychol.*, 52, 175.

HORST, P. (1959). Differential prediction of academic success. *Report*, Office of Naval Research Contract 477 (08) : Univ. of Washington.

HUDSON, L. (1960). A differential test of arts/science aptitude. *Nature.* London, 186, 413.

HUDSON, L. (1961). Arts/science specialisation : a study of differential aptitude. *Unpublished Ph.D. Thesis* : Cambridge University.

PETRIE, A. (1948). The selection of medical students. *Lancet*, 2, 325.

RAPAPORT, D. (1957). " Cognitive Structures." In *Cognition : the Colorado Symposium.* Cambridge, Mass. : Harvard Univ. Press.

ROE, A. (1953). A psychological study of eminent psychologists and anthropologists, and a comparison with biological and physical scientists. *Psychol. Monogr.*, 67, No. 352.

SUPER, D. E. (1957). The Multifactor tests : summing up. *Pers. Guid. J.*, 36, 17.

(*Manuscript, received, 27th October,* 1962.)

CHANGES IN TERMAN-MERRILL I.Qs. OF EDUCATIONALLY SUB-NORMAL BOYS

BY

C. S. RUSHTON

(*Sometime Educational Psychologist, Kent School Psychological Service*)

AND

A. E. STOCKWIN

(*Sometime Deputy Headmaster, Seabrook Lodge School, Hythe*)

SUMMARY. Test results on the Terman-Merrill* 1937 revision of the Stanford-Binet Intelligence Scale (Form L) were examined for 111 'educationally subnormal' boys in a residential special school. A comparison between results on entry at the age of 7 years and above, and before leaving between 14 and 16 years, showed a deterioration in mean I.Q. scores, despite the fact that the Roberts-Mellone† (1952) corrections were applied.

The proportion of 'verbal subtests' passed was consistently lower than the proportion of 'non-verbal' tests,' at all age levels.

The hypothesis that the deterioration in mean I.Q. scores between mean C.A. 9⅙ and 15½ years, was due partially to the increased verbal weighting on this test at the higher age levels, was examined and supported.

It was suggested that the test under consideration is unsuitable for the all-round cognitive assessment of educationally sub-normal‡ children. Suggestions were made for a more suitable type of test, and for further research on Terman-Merrill (L).

I.—INTRODUCTION AND PREVIOUS RESEARCH.

PREVIOUS investigations on this subject have yielded conflicting results. Many of the earlier findings were consistent in that the mean I.Q. of dull children appeared to decline with age. Thus, Phillips (1940) concluded that for children with Terman-Merrill I.Qs. between 50 and 80, there was a tendency for a general and substantial decline which increased progressively with age. Parker (1934), Thorndike (1940) and Stott, (1960) all show increasing deterioration of I.Q. scores with increasing interval between test and retest.

Perhaps the first and most striking contradiction of these findings came from Wellman (1934), who reported that all the groups which she tested tended to show improved I.Qs. while 'the most backward group' showed the greatest increase. Schmidt (1946) and the Clarkes (1958 and 1959) also reported striking gains in the I.Qs. of the 'feeble-minded,' but the former study (Schmidt's) was severely criticized by Kirk (1948) and cannot now be taken seriously.

More recent studies have had available more accurate measuring techniques. Collman and Newlyn (1958) for instance, on a test-retest study of educationally subnormal pupils in special schools in Essex, found that the use of the R.M. correction table (1952) was fully justified. On re-test they found no significant mean change in I.Q.—but this was after a period of only one year. Hiskey and Sadnavitch (1958) and Scarr (1953) report similar findings.

---

* From this point onwards referred to as T-M (L).

† From this point onwards referred to as R.M. corrections.

‡ From this point onwards referred to as E.S.N. children.

In a private communication, P. F. Portwood states that when I.Q. scores on T-M (L) were considered without the R.M. corrections being applied, there appeared to be a significant drop in I.Q. scores after a two-year interval. When the corrections were applied, however, the significant drop (due to the increasing S.Ds. between these age levels and thus to a spurious statistical effect) was changed to a slight though statistically insignificant increase. Here again is evidence which appears to confirm empirically the validity of the R.M. corrections, and the advisability of using them when dealing with T-M (L) I.Qs. on E.S.N. pupils.

II.—Present investigation—method.

A comparison was made between the T-M (L) I.Qs. of 111 E.S.N. boys in a residential special school in Kent on 'First Testing' (usually at the age of 8 or 9 years), and 'Last Testing' (between the ages of 14 and 16 years). All testing was done by assistant school medical officers or by the deputy headmaster, who had been trained in the administratisn of T-M (L).

For testings below C.A. 15 years, R.M. corrections were applied to all I.Qs. Above the age of 14 years 11 months no R.M. corrections are published, yet for the 15-year-old age group the mean I.Q. is 102·8 and S.D. 19 (Terman and Merrill 1937), and thus needs correction. MacNemar, in his earlier correction table (1942) used the same corrections for C.As. 14y. 6m. to 15y. 5m., as Roberts Mellone do for C.As. 14y. 3m. to 14y. 5m. (1952). Thus, it was considered justifiable to use these corrections for test results in the present investigation between C.A. 15y. 0m. and 15y. 5m. Above this age MacNemar does not offer a correction, and at the 16-year age-level S.D. drops to 16·5 (Terman and Merrill, 1937). Thus, for all test results above C.A. 15y. 5m., no correction was made.

Means and S.Ds. were computed on these I.Qs., and significances of differences between means were obtained, using Sandler's 'A Test' (Sandler 1955), which he claims has been 'rigorously derived' from Students '*t*' Test, and corresponds to the two-tail form of '*t*.'

III.—Results and their treatment.

A. *Subjects*. Of the 111 boys studied :

(1) 69 per cent. came from homes in 'urban' areas, and 31 per cent. from 'rural' areas.

(2) 71 per cent. came from families where the father was 'unskilled' or 'semi-skilled' (Grades IV and V of the Registrar-General's categories of occupation groups), as contrasted with 26½ per cent. in the general population of Kent (H.M.S.O. 1954). This finding is in line with the results of earlier researches. (Jones 1954).

(3) Only 29 per cent. came from families in the 'higher' occupation groups (Registrar-General Grades I, II and III), as contrasted with 73·4 per cent. in the general Kent population (H.M.S.O. 1954).

(4) 62 per cent. came from families where there were three or more children, as contrasted with the normal 27 per cent. of families in Kent of this size (H.M.S.O. 1954). On the other hand, only 14 per cent. of our E.S.N. population were 'only' children, as opposed to 42 per cent. in the general Kent population (H.M.S.O. 1954). Eysenck (1953) also reports that dull children tend to come from large families.

B. *Test Results.*

(1) For the population of E.S.N. boys tested, there is a statistically significant 'deterioration' in mean T-M(L) I.Qs. between the ages of 9y. 2m. (before entry to residential special school) and 15y. 6m. (just before leaving special school). (See Table 1).

TABLE 1

TEST-RETEST DATA T-M (L) I.QS. WITH R.M. CORRECTIONS APPLIED.

(N=111)	First Testing	Last Testing
I.Qs.: Means S.Ds. Range	 71·0 7·91 47 to 96	 66·0 9·43 46 to 93
C.As.: Means S.Ds. Range	 9y. 2m. 1·72y. 6y. 11m. to 13y. 4m.	 15y. 5½m. 0·67y. 14y. 0m. to 16y. 1m.
Difference between I.Qs. on first and last testing	5·0 I.Q. points	
Sandler's A	·0308	
Level of Significance	Below ·001	
Product Moment "r" of I.Qs.	+0·626	

TABLE 2

ANALYSIS OF I.Q. CHANGES ON T-M (L) WITH R-M CORRECTIONS APPLIED.

	I.Q. Deteriorated more than 3 points		I.Q. Constant ±3 points		I.Q. Improved more than 3 points		Chi²	Level of Significance
	N	%	N	%	N	%		
Total N=111	68	61%	27	24½%	16	14½%	4·41	P ·05
First I.Qs.	N		N		N			
Below 60 61 to 70	1 22		5 8		7 3		4·59	P ·05
Below 60 71 to 80	1 39		5 10		7 5		5·06	P ·05
Below 60 81+	1 6		5 4		7 1		3·17	P at ·07

(2) The correlation coefficient ($r = +0{\cdot}626$) is in general agreement with the figure which one would expect after a test-retest interval of about 6½ years. Eysenck (1953) and Stott (1960) report that the evidence from thirty similar test-retest studies indicates a correlation of ·79 after 4 years, with a drop of about ·04 points in the coefficient for every additional year intervening. On these predictions, for a test-retest interval of 6½ years, one would expect 'r' to be +0·69., whereas we find +0·626. This lower correlation may well be due to the fact that many different school medical officers were involved in the testing.

(3) From Table 2, it is clear that many more individuals show 'deterioration' than 'improvement' in I.Q. on final testing. Chi^2 test, where 'I.Qs. constant' were apportioned equally to 'improved' and 'deteriorated,' showed that this greater number of deteriorating scores was significant at a ·05 level.

(4) If one sub-divides the data, however, into I.Q. ranges, it can be seen that there appears to be a 'cut-off point' around I.Q. 60 on first testing. Those boys with an initial I.Q. of below 60 on first testing, tend to show an improved I.Q. (more than 3 points) on final testing. Those boys with an I.Q. of more than 60 on first testing tend to show a deterioration of I.Q. on final testing of more than 3 I.Q. points.

These differences in the direction of I.Q. change appear to be statistically significant, and greater than one would expect by chance alone. It may be that they are partially explicable by the concept of 'statistical regression towards the mean" (McNemar 1940 and Clarke, Clarke and Reiman 1958), since they occur in different directions in the two extreme groups, and *towards* the group means. The fact that the group mean itself changes between first and last testing cannot, however, be explained on these grounds, as it changes in the 'wrong' direction, i.e., *away* from the mean score of the population.

IV.—EXPLANATORY HYPOTHESIS.

A.—*Procedure and Findings.*

In a private communication, Professor A. D. B. Clarke, suggested that the most likely reason for the decline in scores was that "the test material is more verbally biassed on retest (Mean C.A. 15·5½), than on initial test" (Mean C.A. 9y. 2m.). His argument is based on the hypothesis that "the increased verbal weighting at the higher age levels is responsible for decreasing the final I.Q., since commonly verbal assets are more seriously impaired in E.S.N. children than are performance assets." (See also Tansley and Gulliford 1960).

In order to test this hypothesis, the sub-tests of T-M (L) were classed as either 'Verbal' (V) or 'Non-Verbal' (including 'Arithmetical') (P) on the basis of the classification of test responses published as Appendix I in Kennedy-Fraser (1948). The original test sheets of fifty-one boys which were immediately available were then re-examined, and each sub-test attempted at either first or final testing was labelled 'V test' or 'P test.' Totals and Chi 2 of their differences are shown in Table 3.

B.—*Indications and discussion.*

(1) On Verbal Tests there is no statistically significant difference between success rates on first and final testings. The direction of the change, however, is towards an improvement in 'verbal efficiency.'

(2) On Non-Verbal Tests there is no statistically significant difference between success rates on first and final testings. The direction of the change, however, is towards a deterioration in 'non-verbal efficiency.'

TABLE 3

DIFFERENCES BETWEEN VERBAL AND NON-VERBAL TESTS ON T-M (L) WITH R-M CORRECTIONS APPLIED.

(N Boys=51)

	VERBAL TESTS			
	Passed	Attempted	Percentage Passed	Chi^2
First Testing	352	927	38%	0·865 P=·40 NOT significant
Last Testing	689	1580	43·6%	
	NON-VERBAL TESTS			
	Passed	Attempted	Percentage Passed	Chi^2
First Testing	349	519	67·3%	2·80 P=·10 NOT significant
Last Testing	192	328	58·5%	
	FIRST TESTING			
	Passed	Attempted	Percentage Passed	Chi^2
Verbal Tests	352	927	38%	58·2 P=less than ·001 Highly Significant
Non-Verbal Tests	349	519	67·3%	
	FINAL TESTING			
	Passed	Attempted	Percentage Passed	Chi^2
Verbal Tests	689	1580	43·6%	12·71 P=less than ·001 Highly Significant
Non-Verbal Tests	192	328	58·5%	

(3) On both first and final testing there is a consistent and highly significant difference between success rate on Non-Verbal tests and on Verbal Tests. E.S.N. boys do better on P tests than on V tests from ages 8 and 9 years up to ages 15 and 16 years. Generally speaking, they passed only four out of 10 V tests, but six out of ten P tests.

Is Clarke's hypothesis that " the increased verbal weighting at higher age levels is responsible for decreasing the final I.Q. " supported?

The experimental data show that on first testing, out of a total of 1,446 sub-tests attempted, 64 per cent. were V tests, while only 36 per cent. were P tests. On final testing, however, out of a total of 1,908 sub-tests attempted, 83 per cent. were V tests, and only 17 per cent. P tests. Since there is no statistically significant difference on V tests between success rate on first testing and on final testing, let us presume that there is no difference at all. We may assume that on V tests the success rate is 40·8 per cent. (mean of 38 per cent. and 43·6 per cent., the actual success rates achieved). Similarly for P tests we assume a success rate of 62·9 per cent. (mean of 67·3 per cent. and 58·5 per cent., the actual success rates obtained).

Let us next presume that on *first* testing 100 tests were attempted—of these sixty-four would be V tests and thirty-six P tests. Applying the differential success rates just calculated, this would mean that twenty-six V tests would be passed, and 22·6 P tests would be passed. Total tests passed on first testing would thus be 48·6.

If we then presume that on final testing, 100 tests were attempted, of which eighty-three would be V tests and seventeen would be P tests, applying the differential success rates again, this would result in thirty-four V tests being passed, and 10·7 P tests. The total tests passed on final testing would, therefore, be 44·7 as opposed to 48·6 on first testing.

Thus, if the actual number of tests attempted on first and last testing was held constant at 100, then considering the different proportion of V and P tests attempted, there would, on balance, be an apparent deterioration in the numbers of sub-tests passed, given no change at all in ' efficiency level ' (i.e., in success-rate) from one testing to the other.

In this hypothetical case then, it would appear that Clarke's hypothesis is supported, i.e., that the increased verbal weighting on final testing is responsible for decreasing the final score.

If we now consider the actual ' raw scores ' from the experiment, we find the success rates for V tests and P tests combined, to be 48·5 per cent. on first testing, and 46·2 per cent. on final testing. Chi^2 of this difference is 0·93, which is not statistically significant.

These percentages are similar to the proportions passed in the hypothetical case. The fact that in the real experiment the percentage of all sub-tests passed on final testing is higher than that in the hypothetical case, must be due to the slightly improved ' verbal efficiency ' on final testing. (See Table 3). 83 per cent. of all the sub-tests on final testing were verbal and the improved ' verbal efficiency ' on these outweighs the influence of the deteriorated ' non-verbal efficiency,' on the remaining 17 per cent. of tests taken.

C.—*General discussion.*

This ' over-weighting ' of T-M (L) with verbal sub-tests from Yr. VII level and upwards, has been criticized elsewhere (Clarke and Clarke 1958 and *Bull. Brit. Psychol. Soc.* 1958). The results of the present investigation would tend to supoprt these criticisms, though not at an acceptable level of statistical significance. The fact that T-M (L) ' markedly penalizes ' the dull person, especially if he is from ' poor or adverse surroundings,' would make it an unsuitable measuring instrument for use in the testing of E.S.N. children, since although it ' indicates his verbal handicaps ' effectively, it fails to give us an all-round assessment of his wider intellectual capacities.

In the present study, Roberts-Mellone corrections were used throughout, a procedure which has been confirmed empirically since the publication of their original paper (1952), following MacNemar (1942), by Collman and Newlyn (1958), by Hiskey and Sadnavitch (1958), by Scarr (1953) and by Portwood (private communication). Such evidence would appear to outweigh the arguments of Stott (1960), which are not always clear, neither are his basic assumptions universally acceptable.

Some of Stott's findings (e.g., greater deterioration in the scores of girls than of boys) suggest that the consistent deterioration which he finds may be due partially "to real changes in the rate of mental development" of such children, yet it is notable that he concludes, "this is not to deny that part of the variation is also due to faulty standardization." Present results would suggest that the deterioration is partially due to the actual structure of the test, which varies in its composition from one mental age level to another, and tends to lower the scores of those children with comparatively poor verbal ability, in direct proportion to the verbal weighting in the selection of sub-tests used at any one testing.

Stott's (1960) paper fails to supply adequate information about the chronological ages of the children tested, about their mental ages, or about the average re-test interval for the children re-tested after 3 years 5 months. The present study offers data on a more homogeneous population, with a test re-test comparison between scores at junior school age, and then nearing school-leaving age. It is notable that in some of the studies discussed above, whether or not R.M. corrections were applied, there is a greater deterioration in test scores, the greater the interval between test and re-test. In the present study, where the test re-test interval is longer than in most, there is a greater deterioration, which suggests that the same aetiological factors may well be cumulatively operative on all the subnormal populations studied.

Various theoretical issues remain to be considered. The work of the Clarkes (1958a, 1958b and 1959), of Mundy (1957), and of Penrose (1949) strongly suggests that for the sub-group we have been studying, there is a relative constancy of 'mental efficiency' (and thus of I.Q.), during childhood, but that after the age of about 15 years, there are often marked improvements in I.Q. Clarke's (1959) hypothesis is that such increments amongst the 'feeble-minded' represent "not a response to present conditions, but rather a recovery from early deprivation," maternal or environmental. This work by no means contradicts the well-known thesis of Bowlby (1951 and 1956) on deprived children, whose follow-up studies have, unfortunately, as yet all been on children below the age of 14 years; it merely has a bearing on the long-term outcome.

O'Connor (1956) points out that the concept of deprivation is still confused, but that one important variable in the normal development of abilities and character, is a 'stimulating environment.' This point of view is supported by Sontag, Baker and Nelson (1958). They found that amongst pre-school children 'emotional dependence on parents' was associated with deteriorating I.Qs. during that period—such children were establishing behaviour patterns which were not conducive to 'learning to learn' during these early years. On the other hand, pre-school children who developed behaviour patterns characterized by aggressiveness, self-reliance, initiative and competitiveness (those from 'stimulating environments'), were developing personality traits which result during school life in gradually improving I.Qs. Such children seem to develop a high need for achievement, which operates as a motive for learning.

In the present paper it is established that E.S.N. boys from 'working class homes' (i.e., from the two lowest social-occupation groups, from which 71 per cent. of our population were derived), score consistently better on the non-verbal tests in T-M (L) than on the verbal tests. Bernstein (1958 and 1959) found that of 309 G.P.O. messenger boys in adolescence from 'unskilled and semi-skilled backgrounds,' 206 scored Matrices I.Qs. of 101+ while the means of their Mill Hill Vocabulary scores fell below this.

It has been argued by Hollingworth (1920) and others that in subnormal children the curve of 'mental development' rises constantly, but more slowly than in normal, and that there is a cessation of growth at an earlier age than for normal children. If this is so, and the curve of mental development is 'less than linear,' then the I.Q. of subnormal children must inevitably fall. Some support has been given to this hypothesis by earlier investigators such as Phillips (1940). In more recent studies, however, where the varying S.Ds. of T-M (L) have been corrected for, no such deterioration has been found, e.g., Collman and Newlyn (1958), Hiskey and Sadnavitch (1958) and Portwood. We have seen too in the work of the Clarkes that marked rises in I.Q. level may occur at adolescence; and Davis (1957) has challenged the earlier assumption, arguing that environmental influences are extremely important in influencing the course of 'mental development.'

The influence of environment may also be important in determining the significant differences found in the present study between the efficiency of E.S.N. boys on non-verbal and verbal tests. We know that they tend to come from large 'working class' families, where parental training, encouragement and concern tend to be lacking, and where other investigators have found children with poor verbal and academic abilities (Nisbet 1953, Anastasi 1956). Such children, like adolescent delinquents (Glueck and Glueck 1950), tend to do relatively well on 'non-verbal tests' whether they are of a performance or perceptual type. Pringle and Bossio (1958) have presented evidence that groups of deprived children obtain lower Verbal than Performance W.I.S.C. I.Qs. Tansley and Gulliford (1960) report that E.S.N. children in a residential special school tend to score higher W.I.S.C. Performance than Verbal I.Qs. These writers suggest that the differences may relate to the fact that "many of their children come from broken or unstable homes, which do not provide the best conditions for language development." In this connection, Bernstein's hypotheses (1958 and 1959) offer a possible explanation of the underlying causes.

In view of the recent appearance of Form L-M (Terman and Merrill 1960), an appraisal was made of the relevance of the present study to the new form, which seems likely to supercede Form L eventually. As in Appendix 1 of Kennedy-Fraser (1948), and as in the present study for T-M (L), the sub-tests of the new Form L-M were similarly categorized as either 'Verbal' or 'Non-Verbal' (which includes 'Arithmetical'), on the basis of the kind of test responses elicited.

Kennedy-Fraser (1948) reports for T-M (L) the following percentages of 'Non-Verbal' sub-tests (including Arithmetical), for M.A. groups Yrs. II—V, Yrs. VI—VII, and Yrs. VIII—S.A.3, 59 per cent., 19 per cent. and 26 per cent., respectively. For the same three M.A. groups, on Form L-M, the percentages of 'Non-Verbal' sub-tests were found to be 61 per cent., 16 per cent. and 30 per cent., respectively. Thus, as far as the 'Verbal'—'Non-Verbal' dimension in the composition of the new Form L-M is concerned, there would appear to be very little difference with T-M (L), and, in fact, many of the sub-tests are the same. The conclusions reached in the present study would, therefore, appear to apply in general to Form L-M.

V.—CONCLUSIONS AND RECOMMENDATIONS.

(1) E.S.N. boys between 7 and 16 years score better on the Non-Verbal than on the Verbal sub-tests of T-M (L). This is in line with earlier findings on the relative efficiency of intellectually inferior groups on verbal and performance tests—E.S.N., deprived and delinquent children and adolescents. This may well be related to the sociological findings that

(2) E.S.N. boys in this Kent residential school tend to come from homes where the father is unskilled or semi-skilled, i.e., from the lowest two occupation groups ; and from larger families of three or more children.

(3) On this experimental population, there are tentative indications that between the mean ages of 9⅙ years and 15½ years, using T-M (L) tests, ' verbal efficiency ' improves somewhat, while ' non-verbal efficiency ' deteriorates a little. Despite these findings and despite the fact that Roberts Mellone corrections were used throughout, there was a statistically significant decline in I.Q. as measured by T-M (L).

(4) This finding may be explained partially at least by Clarke's hypothesis, that it is due to " the increased verbal weighting at the higher age-levels " (83 per cent. as opposed to 64 per cent.), " since verbal assets are more seriously impaired in such children than are performance assets." This is not to deny, however, that other aetiological factors may be operative in causing such a deterioration. Although an attempt has been made by use of the R.M. correction table to correct for the grosser errors of standardization, yet the factor stressed by Stott (1960), " real changes in the rate of mental development," due possibly to the cumulative effects of a culturally deprivatory environment, or to lowered motivation, or to a lowered level of " mental organization and conceptualization," may also be important.

(5) Whatever the causes, T-M (L) would appear to be a misleading test to use in the assessment of E.S.N. children Although it is a predominantly verbal test, and the proportion of verbal items passed improves slightly between 9⅙ and 15½ years, yet its structure is such that the I.Q. appears to decline. It would appear inappropriate for the dull child from a poor or detrimental environment ; it fails to give an all-round assessment of intellectual capacities ; and the different types of items are inconsistently and unequally distributed through the different mental age levels of the scale.

(6) In view of all the difficulties mentioned in this paper, the test under question would appear to need very careful interpretation when applied to an educationally subnormal population, especially if changes in cognitive level were being considered.

(7) The most efficient type of test for use in the assessment of ' dull ' children would appear to be one where verbal/educational and performance/perceptual items were equally balanced, where separate verbal and performance I.Qs. were obtainable, where all the established ' group factors ' were tapped equally and were measurable quantitatively, and where the different types of items were consistently distributed through all age levels of the scale. (A recent attempt has been made by McCarthy and Kirk (1961) to devise a series of tests which analyze the specific psycho-linguistic abilities and disabilities of subnormal children, so that detailed diagnosis can lead to effective remedial treatment. It is as yet too early to judge the effectiveness of such a differential diagnostic instrument, but the question is considered by Cashdan (1962).)

(8) The problems discussed in this paper might well be elucidated, were it possible for some investigator to do a comparable test-retest study on children at the other extreme of the intelligence range—the ' educationally superior ' population. Valuable confirmatory evidence for the main hypothesis of this paper might also be educed, were some of the earlier studies quoted (Phillips 1940, Thorndike 1940, Scarr 1953, Collman and Newlyn 1958, Hiskey and Sadnavitch 1958, Portwood (private communication) and Stott 1960), subjected to the same sort of statistical analysis.

(9) There would appear to be little difference in the 'Verbal'—'Non-Verbal' composition of T-M (L) (1937), and Form L-M (1960). Thus, the above conclusions and recommendations would equally apply to the new Form L-M, as to Form L.

Acknowledgements.—We wish to express our thanks to the County Education Officer, the Principal School Medical Officer and the Children's Officer of the County of Kent and their staffs, for the facilities and information provided; to the Head Master of the School, and the School Medical Officers who did the testing; to Professor A. D. B. Clarke, Dr. N. France, Mr. Gwynne Jones and Mr. P. Portwood for their help and advice. Not least are our thanks due to the 111 children who were the subjects of the study. Opinions expressed in the paper are those of the authors personally and *not* of the Kent Education Committee, or its officers.

VI.—References.

Anastasi, A. (1956). Intelligence and family size. *Psychol. Bull.*, 53, 187-209.

Bernstein, B. (1958). Some sociological determinants of perception. *Brit. J. Sociol.*, 9, 159-174.

Bernstein, B. (1959). A public language. *Brit. J. Sociol.*, 10, 311-326.

Bowlby, J. (1951). *Maternal Care and Mental Health.* Geneva: W.H.O.

Bowlby, J., Ainsworth, M., Boston, M., and Rosenbluth, D. (1956). The effects of mother-child separation—a follow-ups study. *Brit. J. Med. Psychol.*, 29, 211-248.

Bull. Brit. Psychol. Soc. (1958). Appendix I—the 1937 version of the Stanford-Binet Scale—a critical appraisal, 35, 13-15.

Cashdan, A. (1962). The intellectual powers of subnormal children. *Educ. Res.*, 4, 84-97.

Clarke, A. D. B.. and Clarke, A. M. (Eds.) (1958). *Mental Deficiency—the Changing Outlook.* Methuen.

Clarke, A. D. B., Clarke, A. M. and Reiman, S. (1958). Cognitive and social changes in the feeble-minded—3 further studies. *Brit. J. Psychol.*, 49, 144-157.

Clarke, A. D. B., and Clarke, A. M. (1959). Recovery from the effects of deprivation. *Acta Psychol.*, 16, 137-144.

Coilman, R. D. and Newlyn, D. (1958). Changes in Terman-Merrill I.Qs. of mentally retarded children. *Am. J. Ment. Defic.*, 63.

Davis, D. R. (1957). *An Introduction to Psycho-pathology.* O.U.P.

Eysenck, H. J. (1953). *Uses and Abuses of Psychology.* Penguin.

Glueck, S., and Glueck, E. (1950). *Unravelling Juvenile Delinquency.* New York: Harvard U.P.

Hiskey, M. S., and Sadnavitch, J. M. (1958). Minimizing exaggerated changes in Binet ratings of retarded children. *Exceptional Children*, 25, 16-20.

H.M.S.O. (1954). *The 1951 Census Report for the County of Kent.*

Hollingworth, L. S. (1920). *The Psychology of Sub-Normal Children.* New York.

Jones, H. E. (1954). "The Environment and Mental Development" in Carmichael, L. (Ed.) *Manual of Child Psychology.* New York: Wiley.

Kennedy-Fraser, D. (1948). *The Terman-Merrill Intelligence Scale in Scotland*—U.L.P. Scots. Counc. Res. Educ. Pubs. 23.

Kirk, S. A. (1948). An evaluation of the study by Bernardine G. Schmidt (1946). *Psychol Bull.*, 45, 321-333.

MacNemar, Q. (1940). A critical examination of the University of Iowa studies of environmental influences upon the I.Q. *Psychol. Bull.* 37, 63-92.

MacNemar, Q. (1942). *The Revision of the Stanford-Binet Scale: An Analysis of the Standardization Data.* Boston: Houghton-Mifflin.

McCarthy, J. J. and Kirk, S. A. (1961). *Illinois Test of Psycho-linguistic Abilities.* Univ. Illinois.

Mundy, L. (1957). Environmental influence on intellectual function as measured by intelligence tests. *Brit. J. Med. Psychol.*, 30, 194-201.

Nisbet, J. (1953). Family environment and intelligence. *Eugen. Rev.* 45, 31-40.

O'Connor, N. (1956). The evidence for the permanently disturbing effects of mother-child separation. *Acta Psychologica*, 12, 174.

Parker, H. T. (1934). *The Deviation of Intelligence in Sub-normal Children.* Melbourne U.P.

Penrose, L. S. (1949). *The Biology of Mental Defect.* Sidgewick and Jackson.

Phillips, G. E. (1940). *The Constancy of the I.Q. in Sub-normal Children*—O.U.P. Austr. Conc. Educ. Res. No. 60.

Portwood, P. F. (1958). Progress report of the Sheffield E.S.N. Study. *B. Ps. S. Bulletin* No. 34, 61.

Pringle, M. L. K., and Bossio, V. (1958). A study of deprived children, Part 2—Language Development and Reading Attainments. *Vita Humana I*, 3/4, 142-170.

Roberts, J. A. F., and Mellone, M. (1952). On the adjustments of Terman-NMerrill Intelligence Quotients to secure comparability at different ages. *Brit. J. Psychol.* (*Stat. Sect.*), 5, pp. 65-79.

Sandler, J. (1955). A test of the significance of the difference between the means of correlated measures, based on a simplification of Students *t*. *Brit. J. Psychol.*, 46, 225-226.

Scarr, E. H. (1953). Changes in Terman-Merrill I.Qs., with dull children. *Brit. J. Stat. Psychol.*, 6, 71-76.

Schmidt, B. G. (1946). Changes in personal, social and intellectual behaviour of children originally classified as feeble-minded. *Psychol., Monogr.*, 56, 5.

Sontag, L. W., Baker, C. T., and Nelson, V. L. (1955). Personality as a determinant of performance. *Am. J. Orthopsychiat.*, 25, 555-562.

Sontag, L. W., Baker, C. T., and Nelson, V. L. (1958). Mental growth and personality development : a longitudinal study. *Soc. Res. Ch. Def. Pubs.* 23, Antioch, Ohio.

Stott, D. H. (1960). Observations on retest discrepancy in mentally sub-normal children. *Brit. J. Ed. Psychol.*, 30, 211-219.

Tansley, A. E., and Gulliford, R. (1960). *The Education of Slow Learning Children.* Routledge and Kegan Paul.

Terman, L. M., and Merrill, M. A. (1937). *Measuring Intelligence.* Harrap.

Terman, L. M., and Merrill, M. A. (1960). *Stanford-Binet Intelligence Scale.* Harrap.

Thorndike, R. L. (1940). Constancy of the I.Q. *Psychol. Bull.*, 37, 167-186.

Vernon, P. E. (Ed.) (1957). *Secondary School Selection.* Methuen.

Wellman, B. L. (1934). Growth in intelligence under differing school environments. *J. Exp. Educ.*, 3, 59-83.

(*Manuscript reeceived* 29*th, March* 1962.)

THE DEVELOPMENT OF CHILDREN'S IDEAS ABOUT COUNTRY AND NATIONALITY

PART II: NATIONAL SYMBOLS AND THEMES

BY GUSTAV JAHODA
(University of Glasgow)

SUMMARY. The learning of conventional national symbols (national anthem, songs, flags) and the understanding of their significance at various stages were studied. A specially devised picture test was used to explore more subtle aspects of national consciousness. The broad developmental trends and the socio-economic group differences emerging from both parts of the investigation are outlined, and some educational implications suggested.

I.—INTRODUCTION.

IN the literature of the social sciences a good deal of lip-service is paid to the importance of national symbols in the building up of national sentiments. For instance, G. W. Allport (1950) wrote as follows: "As a rule, personal loyalty can adhere to an abstraction only when the abstraction is richly symbolized. Christianity rivets attention upon the cross, nations focus on their respective flags." Few attempts seem to have been made to trace the acquisition of such symbols in young children, apart from the work of Horowitz (1940) and recently Lawson (1963), who concentrated on national flags. The aim of the present study is to initiate the exploration of a somewhat wider range of symbols.

II.—THE NATIONAL ANTHEM.

Six different tunes were played to the children on a tape-recorder in a systematically rotated order. Two of them were Scottish, two Irish, one English, and the National Anthem. They were asked to say of each if they had heard it before; if recognition was claimed, the following questions were asked: Where did you hear it? Do you know what it is? Which country does this tune belong to? Here only results relating to the National Anthem will be given.

Table 1 shows the different kinds of responses. About one-quarter of the youngest children gave no sign of recognition. The second group said they had heard it before, but were either unable to specify the country, or gave absurd answers like "From the country where Herod is" (note the tense), or even "It's from Wagon Train."

TABLE 1
RESPONSES TO HEARING NATIONAL ANTHEM.

	M Schools			*W* Schools		
	6–7	8–9	10–11	6–7	8–9	10–11
Fails to recognize tune	6	—	—	7	1	—
Claims recognition but does not know country, or absurd response	13	4	2	11	6	4
Claims recognition and attributes to Scotland	1	4	2	4	9	9
Ditto and attributes to London or England	3	6	1	—	2	4
Ditto and attributes to Britain	1	10	19	2	6	7
	24	24	24	24	24	24
Per cent. unable to give title of tune	54%	4%	—	87%	12%	—

At this point attention may be drawn to the bottom row of the table, indicating the frequency of ignorance of the name. For this purpose only " God Save the Queen " or " National Anthem " were accepted ; of those rejected, five implied awareness that the tune was linked with the monarchy, such as " The royal song " or " The Queen's music." Beyond the age of 7, inability to name the tune was rare, and in general it was found more often among *W* children (P less than ·02).

Turning now to a consideration of errors of attribution, one finds two major trends. The first, most common with *W* children, was to claim the anthem for Scotland. The explanation which suggests itself is that these children are unlikely to have attained a conception of Britain as a whole, and a check of G-stages supports this : 76 per cent. of them were at or below G.III, as compared with only 30 per cent. of those answering correctly (P less than ·001). The causes of the second error, namely mentioning England or London, are less evident. In some cases at least, judging from the replies, it seems to have arisen from an association in the children's minds between the Queen, her residence in London, and ' her music.'

The pattern of development that emerges from the comments (though owing to limitations of time it was not possible to pursue it systematically) may at least be broadly sketched. At earlier ages, the National Anthem tends to be just one particular piece of music among others. The name is largely unknown, but it is connected with such occasions as school ceremonials, Church and Sunday School (e.g., " We get it at concerts " ; " In the service in the hall—the big ones know the words " ; " It's a hymn ") ; other sources include the wireless and, especially among *W* children, television.* These trends continued at 8–9, but there were some additions. For *M* children, these consisted mainly of Cubs' or Brownies' displays, with *W* ones ' the pictures ' or miscellaneous occasions like " Away up the town when the band was playing." Moreover, the name then being generally known, a definite link with the idea of the Queen had become established. In the case of most of the older (10–11) *M* children, the tune had already taken on the character of a national symbol (e.g., " Well, that can be sung by people all over the British Isles ") ; with the corresponding *W* group, such progress tended to be much slower, as illustrated by the examples below :

> It was on the T.V., when they stand up and sing. (*Why do they stand up* ?) Don't know. (*W*10).
>
> They play it at the start of an international game or something. It could be English, Scottish—anything. (*W*11).

Thus, there is a slow and almost imperceptible transition in the meaning of this music for the children. At first it may be quite unfamiliar or, if recognized, quite wrongly labelled. Later, it comes to be connected first with the person of the Queen, and then with Britain as a whole ; at that stage it was almost invariably correctly named. It was striking to observe how older children, who had already grasped this, responded to the questions with a contemptuous laugh, because they felt the answers were so obvious ; one would hardly have guessed that their insight is likely to have been a fairly recent acquisition

III.—NATIONAL SONGS

Whilst the National Anthem is confined to more formal or ritual occasions, there are many popular tunes tinged with patriotic sentiments Furthermore, by contrast with the National Anthem, such tunes tend to be associated with the various units which together constitute the United Kingdom. The tracing of development in this sphere was the object of putting the following question: Many people from this country go to live in far-away places and are a long way from home. Now imagine they would like to hear three songs: which ones do you think they would choose?

---

* One-third of the 7-year-old *W* children were able to report that " It's on the T.V. when it's finished."

Answers were divided into three categories. The first consisted of those unable to give the name of any song, together with those who produced irrelevant ones. For all these the question effectively meant "Tell me the names of three songs." The youngest children in this category came out mainly with religious songs ("Gentle Jesus," "Jesus loves me," "Away in a manger," etc.). At a later age, irrelevant songs tended to become increasingly secular, and some divergence between *M* and *W* emerged. "I saw Mum kissing Santa Claus" or "Jailhouse Rock" *versus* "Copelia" or "An adult might like opera"). Also included in this category, because their number was relatively small, were what might be called transitional answers. In these, the child seemed to have some inkling of what was required, but was as yet unable to bridge the gap imaginatively by putting himself into the position of someone far from home. Examples of this would be children who suggested such people would wish to hear Italian, Indian or other foreign songs; one boy said: "We might go to France next year, so I think they should be French ones"; in Piaget's terms, this would be a typically egocentric statement. All the responses so far are grouped under ' 'Nil or irrelevant" in Table 2.

TABLE 2

SONGS FROM HOME.

	M Schools			*W* Schools		
	6–7	8–9	10–11	6–7	8–9	10–11
Nil or irrelevant response	20	14	5	21	14	4
Understood and produced none or one	3	4	7	2	7	7
Understood and produced two or three	1	6	12	1	3	13

The second category was composed of children who clearly understood what was required ("They will want to hear a Scottish song," or words to that effect), but either could not think of any Scottish song at all, or were able to name only one. There was a small sprinkling of correct non-Scottish answers such as "Carry me home to blighty," which were, of course, credited. The third category was the same, except that the children knew two or three songs.*

One noteworthy feature of this question is the fact that it is one of the few which failed to yield any significant difference between *M* and *W*. From this, it may perhaps be inferred that the establishment of an effective association between native songs and the idea of the homeland is largely independent of differences in both intelligence and social background. It would seem that conceptually the presence of a global, diffuse and undifferentiated notion of, in this case, Scotland is sufficient for the formation of the link. Thus, none of the

---

* Those most frequently mentioned included: Auld Lang Syne, I belong to Glasgow, Scotland the Brave, Roamin' in the gloamin', Loch Lomond, A Gordon for me, Song of the Clyde, A hundred pipers, Ye banks and braes, My heart's in the Highlands.

G.I. children understood the problem; however, the superiority of G.III over G.II was slight and not significant (P between ·30 and ·20). It is also likely that children in both kinds of environments are about equally exposed to patriotic songs, learning them at roughly the same rate.

IV.—THE UNION JACK.

Nine coloured cards picturing, among others, the Union Jack and the Stars and Stripes, were spread out in front of the children. At the outset they were told to point out those they had never seen before, and these were removed. Next, they were asked the names of the remainder, and the countries to which they belonged. Finally, children unable to comply were informed that some of the flags were Scottish, with the request that they should try and pick them out. The last move turned out to have been a mistake, hence, unfortunately, no useful material was obtained regarding the Lion Rampant and St. Andrew's Cross.* The report on the findings will, therefore, be centred on the Union Jack.

Results are set out in Table 3. First of all, there were two 6-year-old *W* children who placed the Union Jack among those they had never seen: they are included in the top row. It will be noted that the children were generally able to name the Union Jack correctly before attributing it to Britain as a whole rather than claiming it for Scotland only. The *M* children were significantly more successful on both counts (P less than ·01).

TABLE 3

THE UNION JACK.

	M Schools			*W* Schools		
	6–7	8–9	10–11	6–7	8–9	10–11
Does not know country or mistakes other than Scotland	5	1	1	15	5	2
Calls it Scottish	14	8	4	4	11	9
Correct	5	15	19	5	8	13
Per cent. knowing the name of flag	24 42%	24 92%	24 96%	24 29%	24 54%	24 83%
Number of children able to identify U.S.A. flag	3	14	22	4	11	15

The youngest children were apt to link the flag with concrete situations in their spontaneous comments, as shown below:

> It's the flag you always put up when the Queen comes. (*M*6).
>
> That's the one you get at the circus—my daddy buys it for me. (*What country*?) It's the Glasgow flag. (*W*6).

The Union Jack was often called "the Queen's flag," indicating an association already seen in connection with the National Anthem. Beyond the age of 7, grosser mistakes were mainly confined to *W* children; apart from calling a variety of different flags " Union Jack," they sometimes used the expression in a completely uncomprehending manner :

---

* This was because of a pervasive tendency to call the whole of the remainder "Scottish," whereby they were bound to score by chance alone.

> I think that's the Jack flag. (*What country* ?) Don't know. (*W*9).
>
> Yes, that's the Union Jack. (*What country* ?) Is it France ? (*W*9)

In contrast, to this, the bottom row of Table 3 shows that *W* children did not do significantly worse than *M* ones in attributing the Star and Stripes to the U.S.A. It might be added that a good many more recognized the flag without being able to specify the country. Here is a revealing instance :

> That's General Custard's flag ; I've seen him on T.V. ; he fights the Indians. (*W*7).

Lastly, a few impressionistic comments will be offered regarding the Scottish emblems. They appeared familiar to many children, though few even among the older children were able to name them correctly. The pervasive influence of television again seems to have made itself felt, as several children mentioned it in connection with the Lion Rampant, this being the symbol of Scottish Television. Some of the *M* children who did best explained that they had learnt it with the Cubs or Brownies. On the whole, there was little evidence that, within the age-range studied, the Scottish emblems functioned as effective symbols of Scottish nationality.

V.—RECOGNITION OF SCOTTISH PICTURES.

This procedure was devised with the object of assessing some of the subtler aspects of children's knowledge of their country. They were presented with sets of three pictures, reproduced in Plates I–V, and required to indicate for each set which—if any—of the pictures is connected with Scotland.

PLATE I—COSTUMES

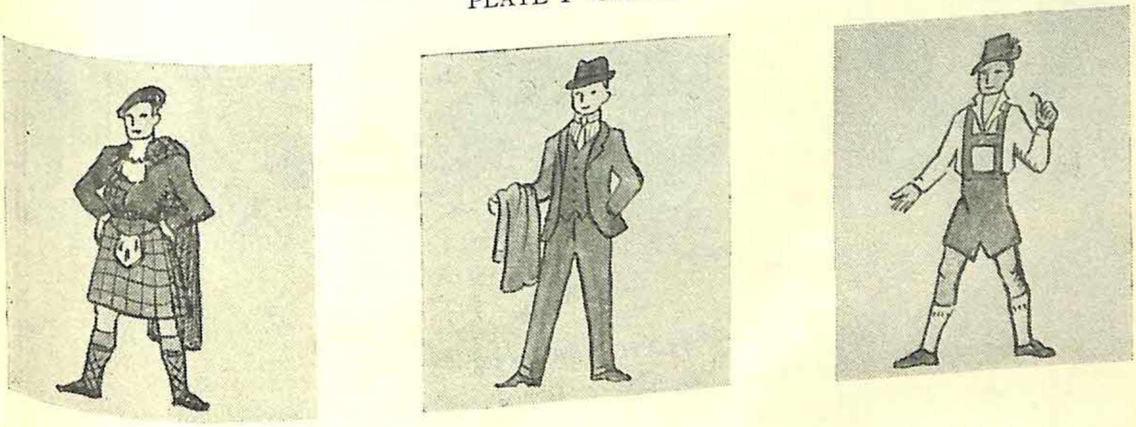


PLATE II—LANDSCAPES

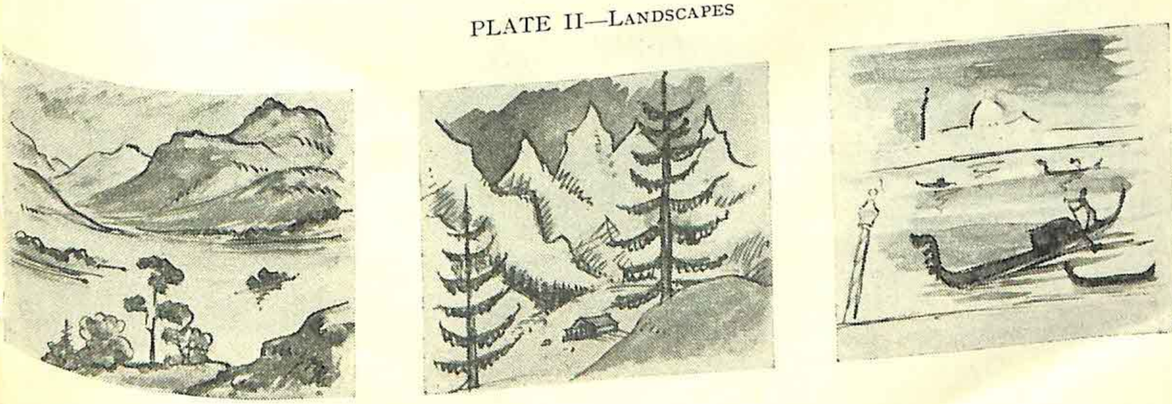
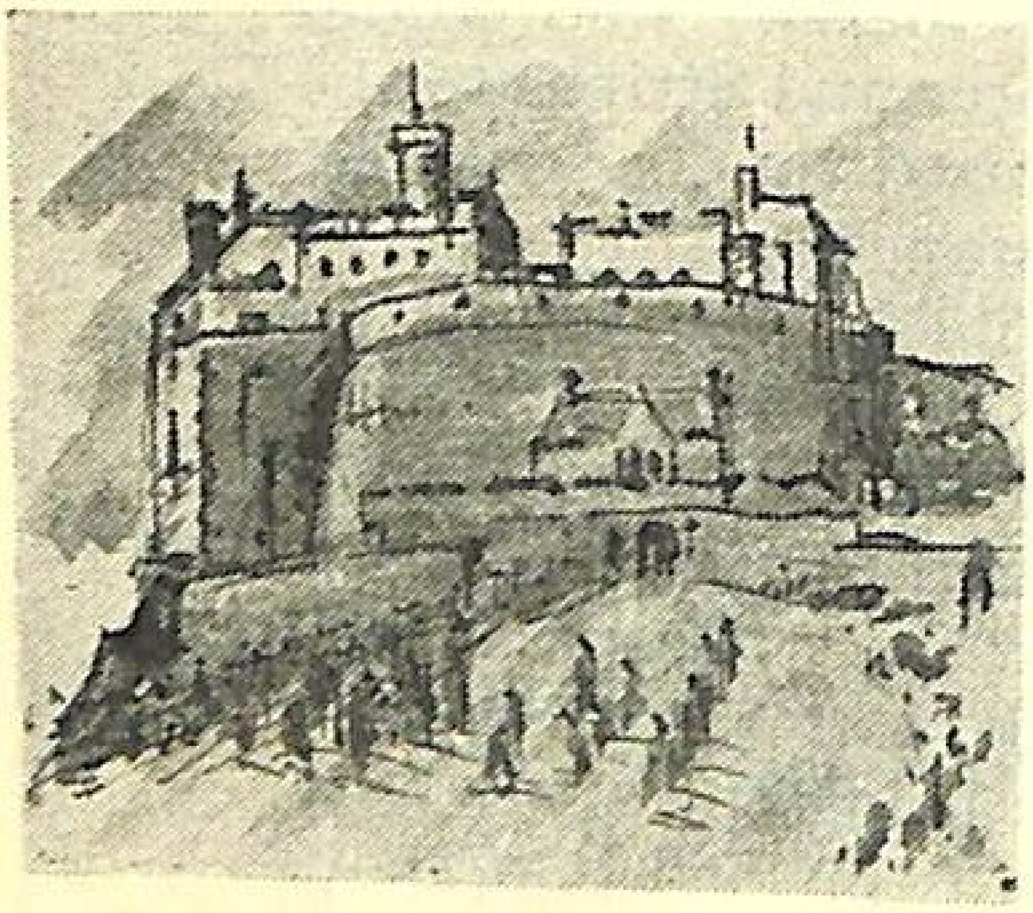
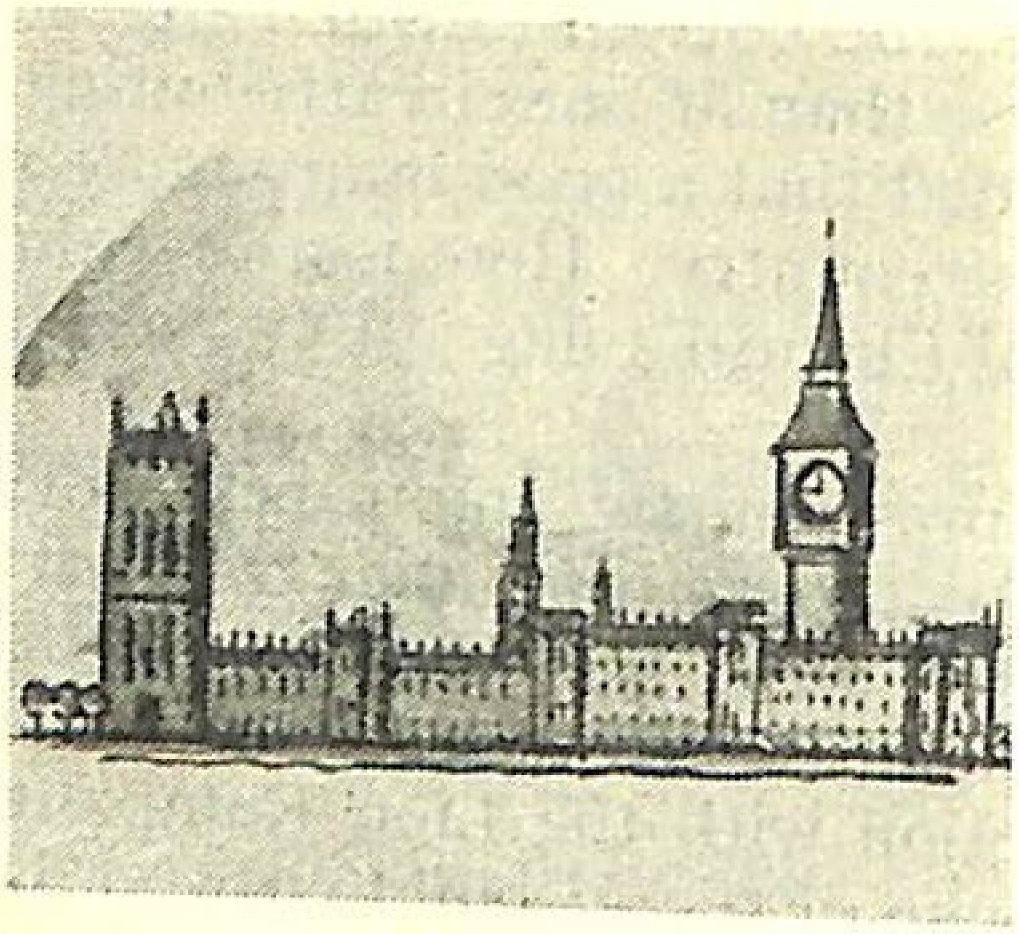


PLATE III—Buildings





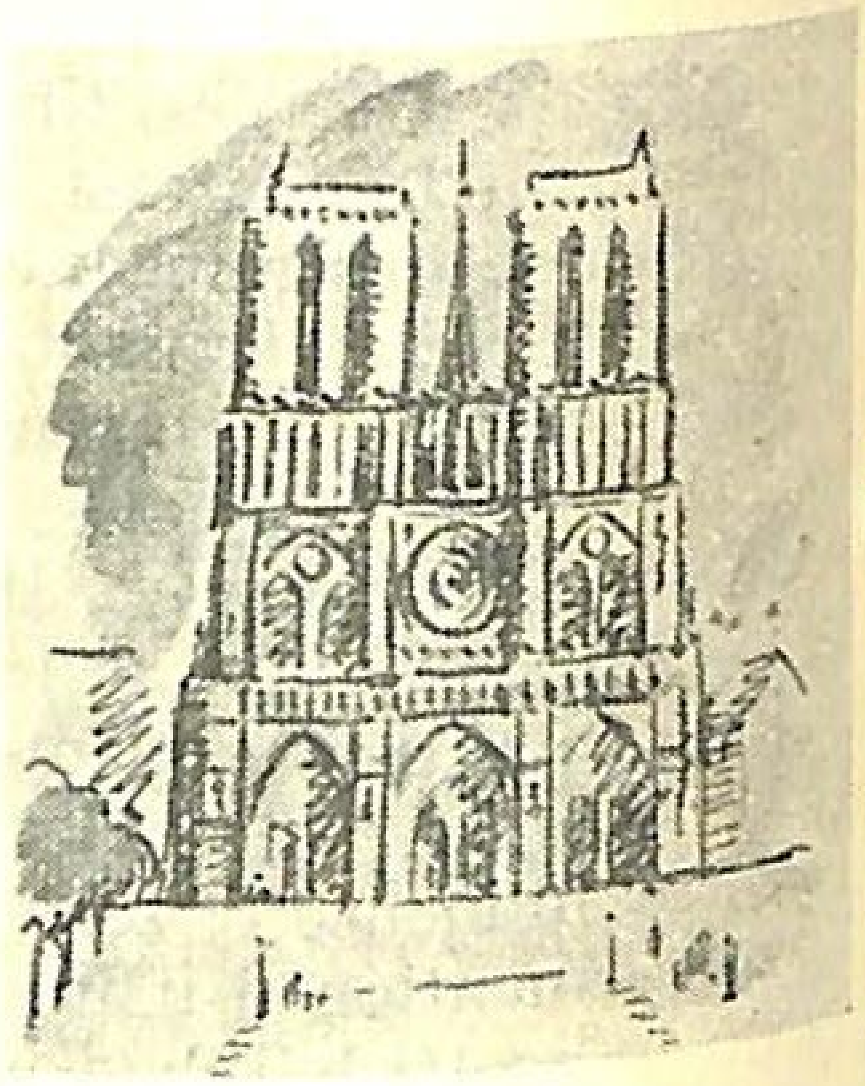
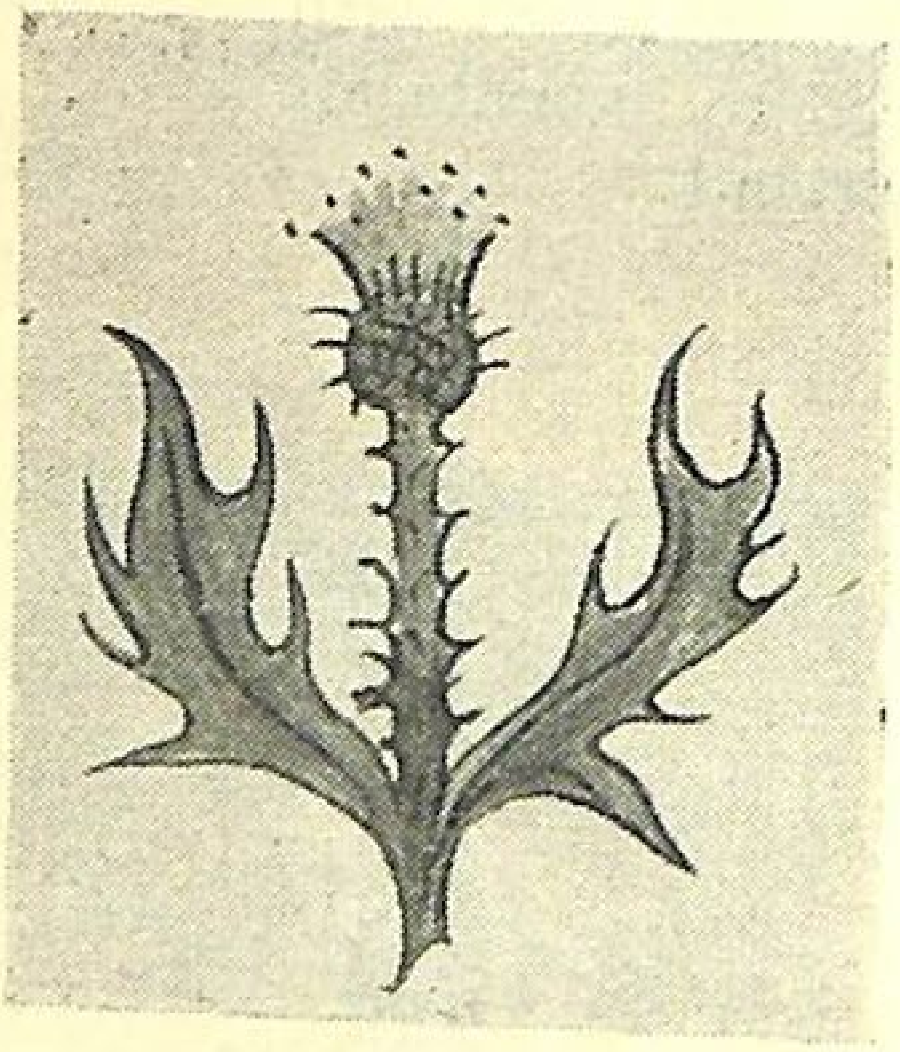
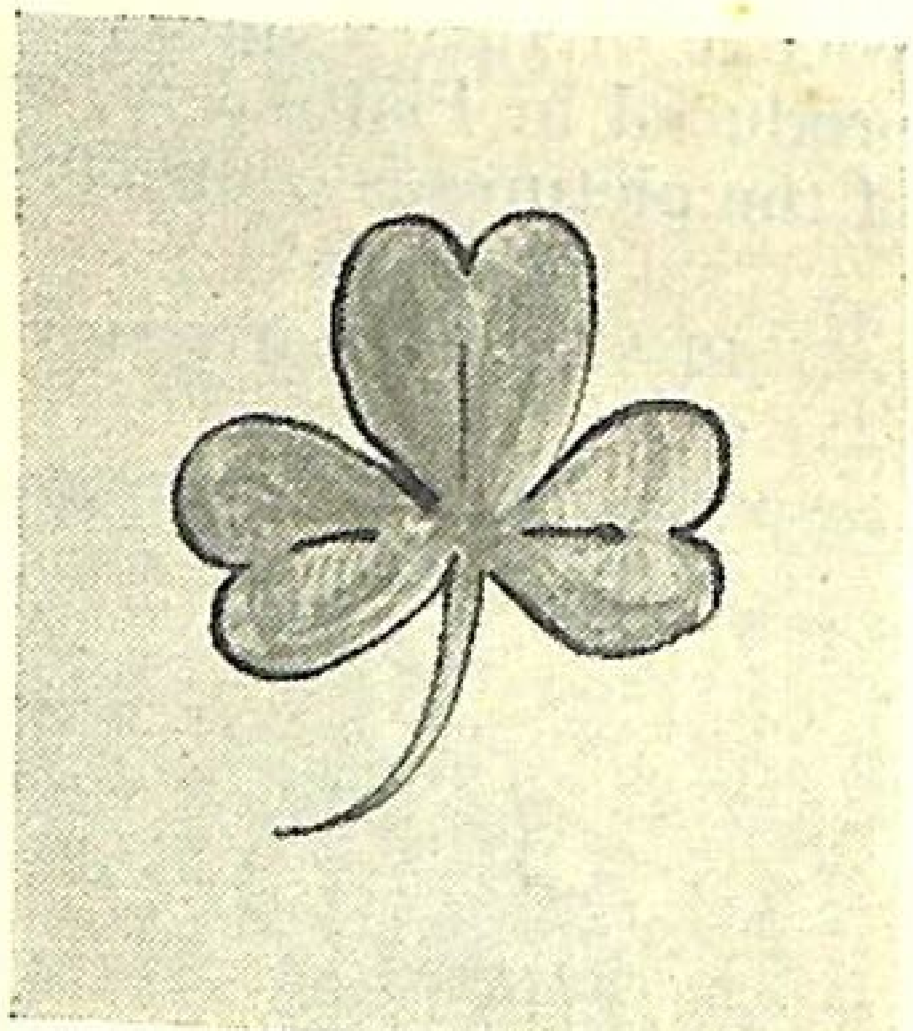


PLATE IV—Emblems





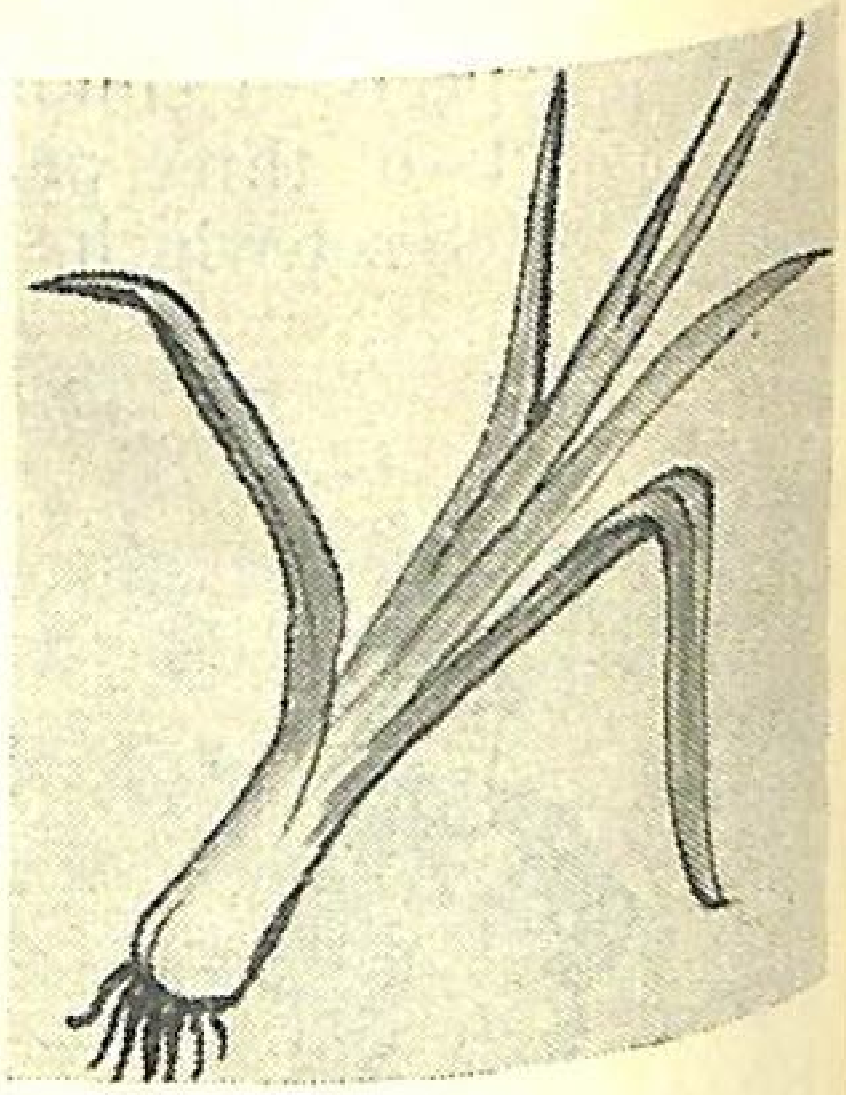
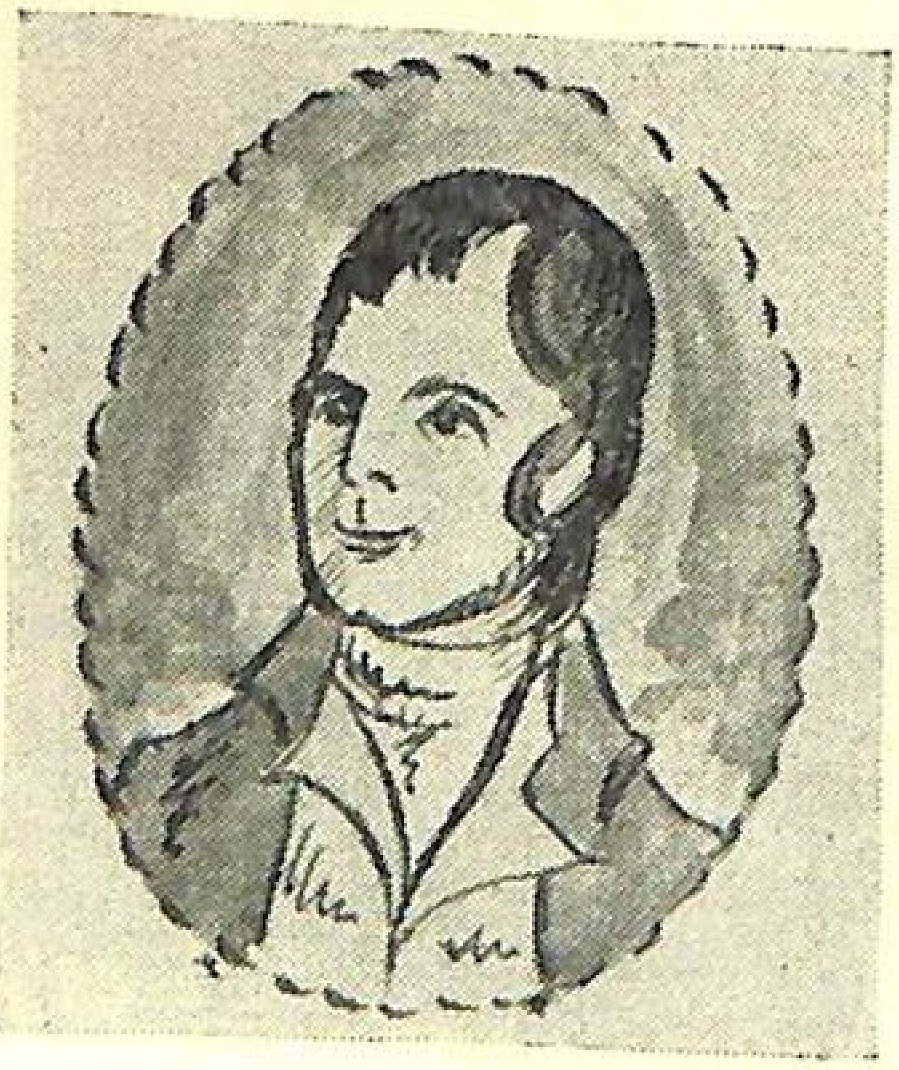
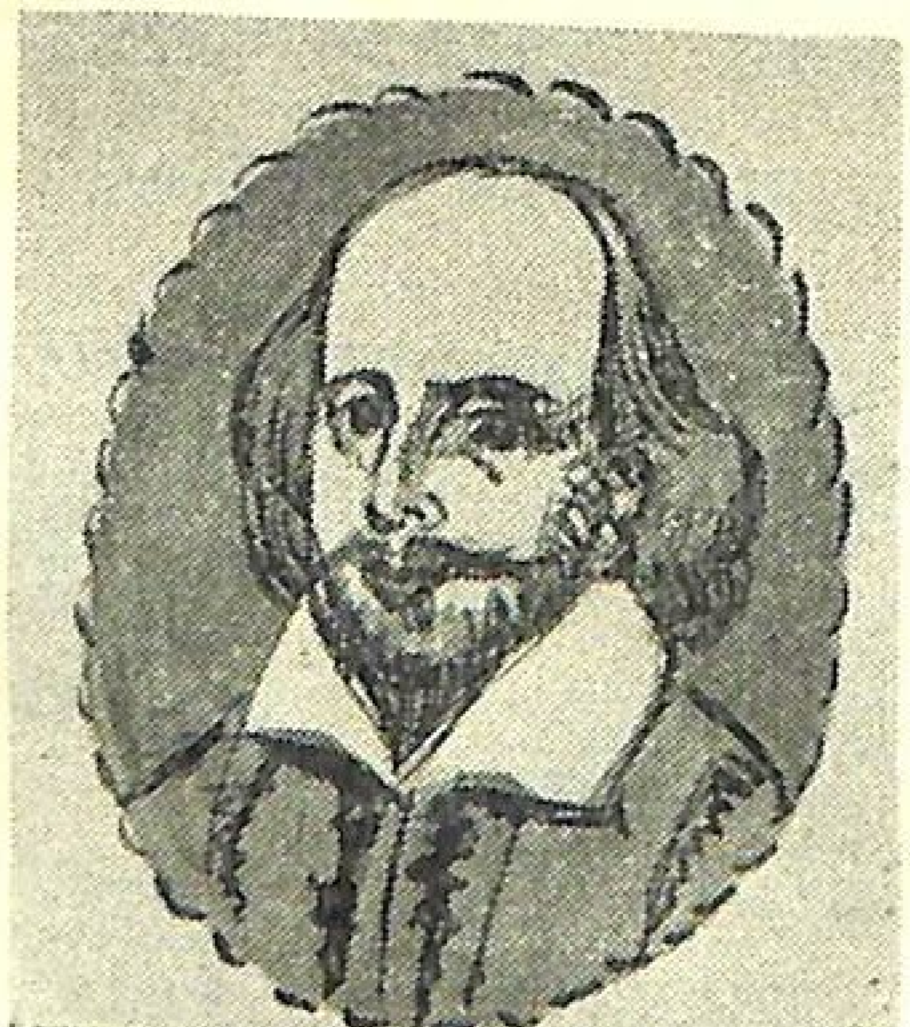


PLATE V—Portraits





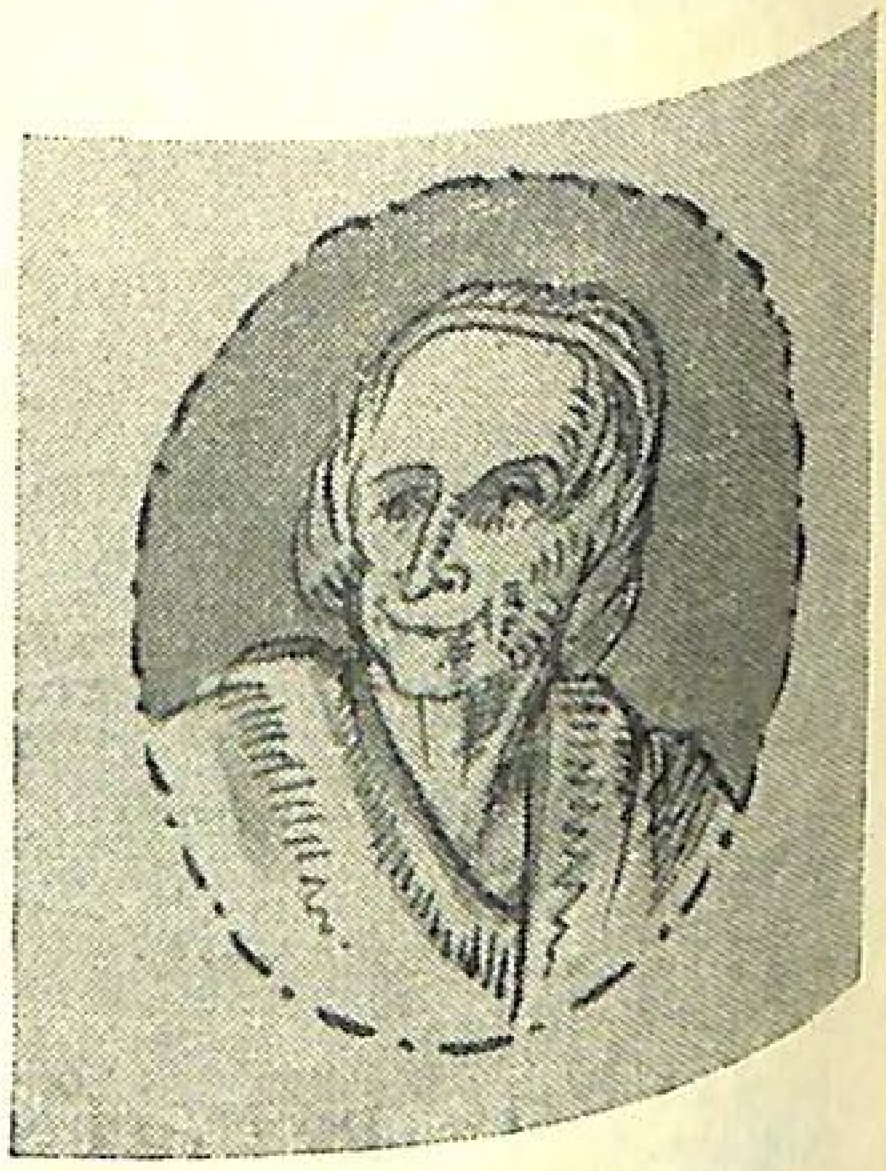


Table 4 gives the results in a somewhat different form, being confined to the number of successes for each sub-group of children. No correction was made for possible chance successes, as it was evident from the children's comments that their number was negligible. The first set, mainly intended as an encouragement, dealt with dress and costumes. The great majority of children had no trouble with this. With the landscape set, where a typical Scottish scene had to be picked out, *W* children were conspicuously less successful (P less than ·001). The difference probably reflects differential opportunities for travelling around the countryside, and possibly also differences in the extent to which landscapes decorate their homes, though the latter is entirely speculative. Similar considerations would apply to the next set, three buildings including Edinburgh Castle (P less than ·02). It is noteworthy that several *M* children were able to identify Notre Dame in Paris, one of the ' distractors.'

After this came a set of conventional emblems, and here the rates of success in singling out the ubiquitous Scottish thistle differed even less, though still significantly (P less than ·05). The last set, featuring " Rabbie " Burns in company with Shakespeare and Voltaire, produced a surprise. There was no significant *M–W* difference, either in correctly choosing Burns as the Scottish one, or even in knowing his name. Evidently Burns constitutes one of the most potent Scottish national symbols, as early as childhood. Below are some revealing answers to the question " How do you know this is Burns ? " :

> My grannie's got a picture of him. (*M*8).
>
> Because we have a calendar with his face on it. (*M*7).
>
> When your mummie's in hospital you go into a fish-and-chip shop for your dinner ; and when I looked up I saw a picture of him looking at me. (*W*6). (N.B.—This boy recognized neither the National Anthem nor the Union Jack.)
>
> On the T.V. news—we saw his cottage. (*W*7).

It would seem that Burns' presence in Scottish life cuts across the barriers of social class, and that is no doubt how he would have wished it to be.

TABLE 4

RECOGNITION OF SCOTTISH PICTURES.

(Numbers of children out of 24 who succeeded on each task.)

	M Schools			*W* Schools		
	6–7	8–9	10–11	6–7	8–9	10–11
Costumes	18	23	24	22	23	23
Landscapes	20	21	23	8	16	18
Buildings	15	21	21	9	15	19
Emblems	18	19	22	11	18	19
Portraits	14	16	22	10	20	22
Percentage of children able to name Burns	29%	46%	88%	21%	58%	88%

VI.—ATTITUDES TO OTHER COUNTRIES AND PEOPLES.

The findings relating to this are reported in detail elsewhere (Jahoda, 1962), and this section, therefore, offers merely a sketch of broad trends and some features of special interest in this country.

The children's growing awareness of more distant nations began with casual glimpses of unusual picturesque detail divorced from any general background. This was sometimes combined with concrete and homely, though equally fragmentary information gained through such personal channels as relatives abroad. The subsequent process of differentiation and organization, too complex for brief summary, was studied in some detail in relation to America and Russia.

Attitudes tended to follow a distinct pattern. Those of the youngest and/or least mature children were determined by some arbitrary and usually concrete detail that happened to have created an impression (e.g., "Don't like New Zealand because the butter is too salt"). Actually, these are probably too evanescent to deserve the term 'attitudes,' and are mentioned only because they are characteristic of an early mode of thought. Apart from the attraction of 'holiday countries,' the more stable sources of favourable attitudes seemed to be again the existence of personal ties, or a belief in the similarity of the people to oneself, resulting in a feeling of affinity; this last statement, it should be added, is somewhat speculative and could not be directly verified from the data. Unfavourable attitudes, from a surprisingly early age, have their roots in past wars, but among the oldest or most mature children these are overshadowed by present political conflicts.

On the salient, and somewhat disturbing findings, was the ethnocentricity exhibited by some *W* children. This took the form of colour prejudice, and a marked tendency to look down on people in other countries. It is interesting to note, however, that their attitude to the English appeared to be influenced by considerations of social class. Thus, a higher proportion of *W* as compared with *M* children thought the English 'nicer' in behaviour (e.g., "In London the men are not so rowdy when they're drunk"); moreover, they tended to identify an English accent with middle class speech, and to regard the Scottish tongue as a working class one, as may be illustrated by two examples:

> Well, the English say "Good jolly, this is a jolly fine day." The Scottish say "Aye, it's a grand day." The English are a lot of toffs. (*W*9).
>
> They talk different—kind of polite, kind of Cockney . . . (*W*11).

Lastly, a few general remarks on attitudes towards the English. From the way the children talked about them (and it must be remembered that will largely reflect the expressed opinion of their elders) it would seem that the traditional antagonism is very much on the wane. There were hardly any disparaging remarks about the "Sassenachs"*, and when questioned about dislikes, not a single one mentioned England or the English.

VII.—CONCLUSIONS.

The outstanding impression one gains from surveying the children's development, in line with the conclusions of investigators concerned with other areas of children's thinking, is that of the great intellectual distance traversed within the span of a few years. At the outset, their responses reveal an almost complete ignorance about the wider geographical and social world surrounding them; some five years later the outlook of most of them is no longer fundamentally different from that of mature adults. During the intervening years, they have acquired a set of conceptual tools enabling them to organize their environment meaningfully and to make the main distinction conventional in our society.

---

* Asked to give another name for the English, only three *M* children were able to produce this ancient word.

The details of this process have been described in this present study. However, it was necessary to treat the various aspects more or less in isolation, leaving aside the question as to whether there may not be some common thread running through the advances in seemingly divergent spheres. An analysis carried out with the material as a whole indicates that the central element is probably to be found in the extent to which the child has succeeded in differentiating and ordering ever-widening sections of his geographical environment; and this can be at least roughly assessed in terms of the G-stages. It would appear that the particular G-stage reached is an index of the limit to the amount of progress possible elsewhere; children often lag behind, failing to realize their potential achievement, but there is no indication that they can exceed it in genuine understanding, as contrasted with mere parrot recital.

Among the youngest children investigated, nearly nine-tenths had not advanced beyond G.II. It is thus legitimate to generalize about the age-group 6–7, because their outlook was fairly uniform. After the age of 7, the increasing diversity of G-stages precludes any general statement, which would be misleading.

It may be recalled that G.I. denotes a perspective confined to the child's immediate surroundings—home, school and play streets in which they spend their everyday life. With G.II an extension has taken place, involving some apprehension of Glasgow as a unitary whole, to which the child refers himself and of which he feels himself to be a part. As yet, however, no conception has been formed about what lies outside. From this starting-point, an inventory of the ideas characterising a majority of the children aged 6–7 will be made.

First of all, the meaning of even the simplest geographical words tends to be extremely vague, or altogether mistaken. Thus 'town' is liable to be applied to any aggregate from streets upwards. This reflects not only inadequate command of this particular concept, but at the same time also inability to classify known geographical names in any coherent order; both are really two sides of the same coin. Asked "what is your country?" the dominant answer was "Glasgow"; one replied "I don't live in a country, I live in a close.* Whilst they have heard the *word* "Scotland," and if G.II usually realize that it is a place of some kind, they cannot locate it properly in space; it is thought of as being "away" somewhere. Requests to say something about Britain elicited a collection of absurdities. The most they achieved, with three exceptions, was to put forward notions implying recognition that Britain is some kind of geographical entity, though invariably a remote one. Consequently, symbols relating to Britain as a whole, such as the National Anthem or the Union Jack, were at best attributed to Scotland; a majority of children in G.I failed to recognize them altogether.

When asked whether they were British, a majority of the children in G.I and G.II replied negatively; the same answer was given by nearly all children in G.I to the question as to whether they were Scottish; but two-thirds of those in G.II affirmed that they were Scottish. In view of the evident confusion prevalent even among the latter, this result may appear surprising: why should they be ready to acknowledge being Scottish, when they clearly fall short of understanding the relationship between Glasgow and Scotland? In order to explain this, one has to abandon the exclusive concern with purely cognitive aspects. These children are growing up in a symbolic environment (i.e., that part of their environment consisting of words, signs, etc.) where Scotland and things Scottish are bound to figure rather prominently. Hence, even before they are capable of building up a coherent conceptual framework serving to categorize their geographical surroundings, they will become familiar with such words as "Scotland" and "Scottish," as well as to a lesser extent "Britain" and "British." Moreover, some broad and loose associations, normally with a favourable emotional tinge, will tend to be established between these words and certain items of their experience. This was often apparent in earlier pages: children were reminded by the word "Scotland" of their family, of Highland Dancers, and so forth. An accumulation of such associations leads to a readiness to acknowledge themselves as "Scottish"; but until such time as an adequate concept of "Scotland" and "Scottish" has crystalized, such quasi-identifications remain unstable and apt to be withdrawn when challenged.

---

* Close: strictly an entrance or passage, but extends to the common stairs of a tenement, or the building as a whole.

The purpose of labouring this point at some length is to show again how young children are sometimes able to use these words in a formally correct manner, when more careful probing would reveal that conceptual grasp still eludes, them. In fact, during the first two or three years of their school career, most children have hardly begun to make sense of the wider geographical, social and political setting around them. It will not have escaped the readers' attention that it was necessary to qualify statements about the youngest children, because even they were far from uniform. Thus, one 6-year-old was already in G.IV, whilst at the other end of the scale, two 10-year-olds had not moved beyond G.II. This variability of individuals cannot be stressed too strongly, because it is in danger of being overlooked when one thinks in terms of age trends, and this applies most of all to the intermediate ages. A majority of the youngest children are likely to be in G.I or G.II, and most of the oldest in G.III or G.IV, but those aged 8–9 are distributed over the whole range. This holds true for children from *M* and *W* schools taken separately, as well as for the total sample.

In general, the children from *M* schools were found to be significantly in advance of those from *W* schools; however, when *M* and *W* children were matched for intelligence test score, the difference disappeared. It was, therefore, concluded that this factor accounts for a large part of both individual variations and group differences.

VIII.—Some educational implications.

"He that hath names without ideas, wants meaning in his words, and speaks only empty sounds." It would be entirely in the spirit of John Locke to add that the passage from names as more or less empty sounds to names backed by fully elaborated concepts is the outcome of a gradual learning process. However, it is not always easy to know whether or not a child who has a name speaks more than empty sounds. An adult may readily be led to assume that notions self-evident to him will also be within the reach of the child. One might again turn to Locke for an example.: "The relation also that things have to one another in their *places* and distances is very obvious to observe; as above, below, a mile distant from Charing Cross, in England, in London." An analysis of the concepts underlying such statements, along the lines sketched in the present study, will indicate that to a child such relationships may be far from obvious.

What bearing does this have on the teaching situation ? The simple answer would be that the teacher should not jump too far ahead of the children's capacity for adequate comprehension ; but this is hardly a very helpful recommendation. In practice, the teacher could not hope to ascertain the stage of conceptual development reached by each child, and even if this could be done, some compromise would be unavoidable. More effective action would be a reconsideration of the manner of teaching such subjects as Geography and History. An examination of the Schemes of Work for teaching these in primary schools, keeping in mind the findings reported, will reveal the presence of numerous assumptions that are likely to be invalid ; for instance, what kinds of 'simple geographical terms' are regarded as having to be taught, and which ones seem to be tacitly taken for granted ?

Another possible suggestion is that some account be taken in planning the curriculum of what appears to the dominant, if not universal sequence of concept development from the immediate vicinity outwards. This would rule out, say, a leap from plans of the school room and adjacent streets to the globe or map of the world. The main aim would be a consolidation of the series widening out from the home town to Britain as a whole.

Finally, it is important to guard against a possible misconception, arising from the general discussion of developmental stages. It has been stated that a child must attain a certain level of intellectual maturity before he is ready to move from one stage to the next ; he must be capable of performing the appropriate logical operations as well as of handling spatial relationships. Thus, subject to the wide range of individual variation already emphasized, there is probably a ceiling to what a given child can understand at a particular time. However, it might be wrongly inferred from this that there is relatively little the teacher can do to help a child along. Such a view presupposes that most children more or less automatically advance to the limits of their potentialities, and this is, unfortunately, very far from the truth. Even in the course of the actual investigation, as an unintended consequence of merely trying to answer questions of a kind with which they were not normally faced, some children had a flash of insight ; things suddenly fell into place, and they reached a new level of understanding. If this can be brought about so casually and incidentally, it is likely that teachers could achieve far more by a systematic effort. It is hoped that the present investigation may make some limited contribution by pointing the way.

ACKNOWLEDGEMENTS.—Thanks are due to Dr. H. Stewart Mackintosh, Director of Education, and to Heads of Schools for their help with the research, and to Mrs. Diana Buchanan who prepared the material for the picture test. Most of all, I am indebted to my colleague Alistair Chalmers who carried out the main fieldwork and helped with the analysis. The research was supported by a grant from the Court of Glasgow University.

IX.—REFERENCES.

ALLPORT, G. W. (1950). In T. H. Pear (Ed.) : *Psychological Factors in Peace and War.* London : Hutchison, p. 153.

HOROWITZ, L. (1940). Some aspects of the development of patriotism in children. *Sociometry*, 3, 329-341.

JAHODA, G. (1962). Development of Scottish children's ideas and attitudes about other countries. *J. Soc. Psychol.*, 58, 91-108.

LAWSON, E. D. (1963). Development of patriotism in children—a second look. *J. Psychol.*, 55, 279-286.

(*Manuscript received 20th August*, 1962.)

THE SELF CONCEPTS OF TRAINEE-TEACHERS IN TWO SUB-CULTURES

By A. S. PHILLIPS

(*Department of Education, University of the West Indies*)

SUMMARY. (1) The aim of this study was to compare the Self Concepts of training college students from two different sub-cultures, viz., England and Jamaica.

(2) Unstructured compositions by students from both countries, on the topic—"I myself: the person I am; the person others think I am; the person I would like to be," were analysed to reveal the categories students use in describing themselves. Some interesting qualitative differences appeared though students from both countries used the same categories in their self description.

(3) Objective tests for measuring the Self Concept of students were constructed and administered to other students in both countries. The following comparisons between the two groups emerged:

(*a*) Jamaican students, both male and female, of the regular age range as well as the mature, rated themselves significantly higher on the Cognized and the Ideal Selves, and were more Self Accepting than their English counterparts;

(*b*) English male students rated themselves at approximately the same level as English female students. But, in the Jamaican samples, the female students tended to rate themselves higher than the men, and to be less Self Accepting.

(*c*) There were no singificant differences between the Self Concept scores of 'regular' English and 'mature' English students. But the 'mature' Jamaican students cognized themselves significantly lower than the 'regular' Jamaican students.

I.—INTRODUCTION.

WITHIN recent years the Self Concept of the individual has been receiving more and more attention from psychologists. It has come to be regarded as the vital part of personality structure, exercising a profound effect on behaviour and on happiness. According to Lecky[4] one of the functions of the Self is to produce an harmonious integrated personality. This is achieved largely by an acceptance of oneself as one is, not in a sense of complacency, but with a full knowledge of the things about oneself one cannot change, and of the endowments one has to work on. Such understanding leads to mental health. Besides, our general behaviour is to a large extent, regulated and organized by what we perceive ourselves to be.

If the individual, therefore, is to lead a reasonably happy, normal, well-adjusted life, free from fears and tensions, from feelings of constant self-reproach and utter inadequacies, it is fairly clear that he should acquire as early as possible, a Self picture that is adequate and realistic, and which will enable him to think of himself in a satisfactory manner. This obviously is a task for the agencies of education. Staines[5] has shown that satisfactory changes can be made in the Self Concepts of children, without any lowering of achievement in the ordinary school subjects, by merely adapting the methods of instruction If the Self Concepts of children are of such vital importance, then the Self Concepts of trainee teachers must be of equal importance and worthy of further study. This research was, therefore, designed to probe the Self Concepts of teachers in training. This was done in two different sub-cultures, viz., Jamaica and England, and comparisons were made between the two groups.

II.—THE JAMAICAN SOCIETY.

The Self Concepts of students from the Jamaican and English sub-cultures might be expected to differ widely as the social forces in these two countries have been so varied. Jamaica has been an English colony for over 300 years. Many races compose the population, but the two dominant ones are the white and the negro. From the beginning of its colonial history, these two races were brought together with very clearly defined places in the social hierarchy. The white people owned all the property and all the wealth and were thus in a position of authority and prestige. The negroes were, in the beginning, merely slaves without rights—the chattel of their masters. Emancipation when it came, secured the physical freedom of the slaves, but the pattern continued. The coloured people were made to feel inferior in so many ways, some subtle, some more brutal, that it was virtually impossible for them to doubt their status. This was strengthened by the processes of education, when education did arrive for the coloured children. Its pattern was copied almost entirely from the United Kingdom. The syllabus of the schools was that designed for the United Kingdom. The standards and the ideals were theirs. Even the illustrations and the pictures presented to children were those prepared for children in England. Only in rare circumstances would a child see a negroid type depicted in an illustration, and even then it was usually conforming to the stereotype of 'backward native.' Even the language they were taught was full of expressions equating black with bad or undesirable, and white with good or desirable. The image held up to the child then, particularly in relation to physical characteristics, would be that of the European, and the European type would certainly form the Ideal Self picture.

This probably explains the coloured person's rejection of the negroid part of his ancestry (there has been much intermixture between the races in Jamaica) and his desire to escape from it, as typified by such practices as straightening the hair and using bleaching creams to lighten the complexion.

Though the more obvious forms of discrimination no longer obtain in Jamaica and there is fairly general equality of opportunity for all races, nevertheless, the attitudes still persist, and might be expected to produce Self Concepts full of unworthiness and inadequacy and marked by excessive rejection of Self. There should, therefore, be marked contrasts with the Self Concepts of training college students from England, where no such forces appear to operate.

III.—SOME UNSTRUCTURED COMPOSITIONS.

As a first step, groups of training college students in England and Jamaica were asked to write unstructured compositions on the topic: "I myself: the person I am; the person others think I am, and the person I would like to be." These three sub-divisions of the main topic were taken as corresponding to the three levels of the Self, viz., the Cognized Self, the Other Self, and the Ideal Self. The aim was to discover the categories that students use in reporting on the Self Concept at these three levels, and also to provide qualitative evidence as to the differences in the way students from both countries cognize themselves.

There was evidence in the essays that students were quite honest in giving their estimates of themselves, perhaps because they had been assured of anonymity. Remarks such as the following were quite frequent:

> "If this were to be read by any one in this college it would be quite impossible for me to write this."

or

> "I find it difficult for me to describe myself, but I shall try to be perfectly honest."

There was some evidence too, that the exercise tended to make the students somewhat uncomfortable, e.g. :

> " I am, at the moment, not liking putting down all these thoughts about myself, because it injures my pride to see myself in such a bad light "

or

> " This hurts, but this is the truth."

An analysis of the essays showed that students in both sub-cultures used more or less the same categories in defining themselves. These categories, with the percentages from each group that gave replies in each category (for the level of the Cognized Self) are set out in Table 1.*

TABLE 1

PERCENTAGE OF STUDENTS WHO MENTIONED ITEMS IN VARIOUS CATEGORIES AT THE LEVEL OF THE COGNIZED SELF.

Categories	English				Jamaican			
	Regular		Mature		Regular		Mature	
	M	F	M	F	M	F	M	F
N—	60	60	—	27	69	25	—	100
1. Physical Characteristics	35	26·6	—	14·8	21·7	32	—	55
2. Clothes, Dress and Grooming	5	13·3	—	11·1	7·3	8	—	26
3. Material Possessions	11·6	13·3	—	7·4	5·8	4	—	13
4. Home and Family	11·6	43·4	—	25·9	18·8	16	—	60
5. Animals and Pets	0	10	—	7·4	0	0	—	1
6. Intelligence	43·3	20	—	7·4	5·8	8	—	12
7. Personality and Character	76·6	93·3	—	100	84·1	92	—	93
8. Philosophical Category	40	38·3	—	25·9	71	48	—	22
9. Performance Category	73·3	73·3	—	39·3	56·6	64	—	86
10. Social Relationships	75	91·6	—	81·5	62·3	68	—	92
11. Status and Role	8·3	21·6	—	3·7	4·4	4	—	8
12. Wants, Goals and Aspirations	6·6	20	—	7·4	16	8	—	16
13. Aesthetic Life	18·3	31·6	—	25·9	10·2	16	—	24
14. Miscellaneous	0	5	—	0	2·9	4	—	2

Certain qualitative differences in the way in which students in each group described themselves also emerged. For instance, there was a tendency for the English students, especially the male students, to denigrate themselves. For example, one man wrote :

> " I make a hell of a lot of noise ; I talk too much, drink too much, lie too much, but I can't work up a conscience about any of these things."

On the other hand, the Jamaican group tended to be largely uncritical of themselves, except in minor facets of personality. They tended to express themselves in superlatives, and to see themselves through very rosy spectacles, e.g. :

> " (I am) a warm-hearted person (and) have some of the finest qualities of which human beings are capable."

---

* For further details, see Phillips, A. S.—" A Study of the Self Concepts of Selected Groups of Training College Students in their relation to other variables in the Teacher Training Situation." Unpublished Ph.D. Thesis, London University.

> " I am . . . person of more than average ability. I am straightforward in all my dealings, thus causing my character to be beyond reproach."

Certain similarities were evident among the various groups, however. Table 1 shows the high percentage in all groups that gave responses in the categories of 'Personality and Character,' 'Performance,' and 'Social Relationships.' This is in accordance with Jersild's^2 claim that mature people tend to see themselves in terms of inner forces rather than of external qualities. The generally low frequency of responses in such categories as 'Material Possessions,' 'Clothes, Dress and Grooming,' 'Status and Role,' etc., applied to all groups in both sub-cultures.

Much greater emphasis was placed by both groups on the Cognized Self, rather than on the other two levels of the Self.* This may have resulted merely from the fact that the Cognized Self appeared first in the essays and so received far fuller treatment, though many students admitted that they had great difficulty in knowing what other people thought of them. Also, some, particularly among the Jamaicans, claimed that they would not wish to be any different from what they were.

IV.—CONSTRUCTION OF AN OBJECTIVE TEST FOR MEASURING THE SELF CONCEPT.

In order to secure more objective and quantifiable measures of the Self Concept, a test of the self rating type was constructed. This was divided into two parts measuring respectively the Cognized Self and the Ideal Self. It was decided to investigate only those categories at these two levels of the Self, as were actually reported on by the students in their essays. Twelve categories only were investigated, as one, viz., 'Animals and Pets' did not appear to be of sufficient universality to warrant inclusion. Later, as a result of statistical analysis, another category, that of 'Aesthetic Life' was also omitted, leaving eleven categories in the final form of the test.

The individual items in the test consisted of statements made by the students in their essays. The main criteria for selection of these statements were that they had to be applicable to people of different races, living in different countries, to both sexes, and to people of varying ages. Also, items had to be positively framed, that is, indicating possession of a desirable quality. Students were asked to rate themselves on a five-point scale (from 0–4) as far as each item was applicable to them. This gave a score for each item, and as items were grouped into categories, a score for each category. A summation of all items gave a total Self Concept score for the level of the Cognized Self. For the level of the Ideal Self, students were asked to rate themselves on these same items, not as they thought they were, but as they would like to be. The algebraic difference between the scores on the Cognized and the Ideal Selves (the discrepancy score) gave a measure of Self Acceptance.

An unequal number of items was selected for each category as some categories were regarded as of wider conception than others. After the first trial run, it was apparent that certain categories hung together, covering more or less the same area of the Self Concept. It was, therefore, decided to group categories into areas. Evidence for such grouping was provided by correlating scores for various categories with each other. Next, the items forming an area were tested for evidence of scalability according to the Guttman technique, and this revealed that they did, in fact, form a scale. On this basis, five clear areas of the Self Concept were revealed, viz. :

(1) Possessions and Desires.

---

* See Phillips, A. S., *loc. cit.*

(2) Personality, Character and Beliefs.
(3) Social Aspect.
(4) Performance Aspect.
(5) External Appearance.

All further comparisons were done in these areas.

The reliability of the test was determined by a re-test after an interval of three weeks, when a product-moment correlation coefficient was calculated. This gave an overall figure of +·87 for the Cognized Self and +·80 for the Ideal Self. This was regarded as adequate.*

V.—Comparison of students from English and Jamaican colleges.

Table 2 gives the numbers and average age of students in each sample. ' Regular ' students were taken as those between 18 and 24 years ; ' Mature ' students were taken to mean those over 25 years. Unfortunately, it was not possible to obtain large enough samples of mature male students in either country.

TABLE 2

Numbers and Average Age of Samples.

Samples	Numbers	Average Age
Regular English Female	58	19 years 4 months
Regular Jamaican Female	61	20 years 4 months
Regular English Male	51	20 years 8 months
Regular Jamaican Male	47	20 years 11 months
Mature English Female	11	36 years 4 months
Mature Jamaican Female	98	30 years 5 months

In comparing scores for training college students from both sub-cultures, only total self concept scores are given here.† These are set out in Table 3.

TABLE 3

Mean Self Concept Scores.

	Jamaican			English				
	N	M	σ	N	M	σ	C.R.	Sig.
Cognized Self								
Regular Female	57	176·50	12·80	58	137·30	19·30	12·90	·001
Regular Male	43	172·58	10·96	49	143·26	18·70	9·25	·001
Mature Female	88	166·32	17·00	11	143·82	20·80	4·05	·001
Ideal Self								
Regular Female	49	195·26	9·45	58	172·87	12·90	10·02	·001
Regular Male	40	181·75	14·62	48	172·63	17·80	2·63	·01
Mature Female	84	190·61	10·70	11	172·45	15·60	4·70	·001
Self Acceptance								
Regular Female	46	17·36	10·91	58	35·02	14·48	7·13	·001
Regular Male	38	10·16	10·10	48	29·92	14·50	7·40	·001
Mature Female	76	24·25	13·70	11	27·91	13·50	·61	N.S.

---

* For complete test, see Phillips, A. S., *loc. cit.*

† For full comparison, see Phillips, A. S., *loc. cit.*

It had been expected that the Jamaican students would cognize themselves less adequately, that is lower than, their English counterparts, and be less self accepting (i.e., have a higher discrepancy score). It was revealed that the Jamaican students, both male and female, those of the regular age range, as well as the mature female students, rated themselves significantly higher on the Cognized as well as on the Ideal Self. They also had a lower discrepancy score (i.e., they were more self-accepting) though this difference does not reach significance for the mature female group. These trends are generally true for all areas of the selves.

It is not easy to explain the higher rating achieved by the Jamaican student over his English counterpart. Perhaps it has its roots in the awakening national and racial consciousness evident in Jamaica over the last twenty years. During this period has come a more advanced political constitution for the island, and a widening of opportunities for its dark skinned inhabitants. An unofficial colour bar, which operated prior to 1940, has disappeared. Generally speaking, the negro has acquired a new pride in himself. But this does not explain why the Jamaican students, even the much older ones, should cognize themselves so much higher than (instead of more nearly equal to) the English students. Another possible explanation is that defence mechanisms are at work. Many forces have been trying to impress on the Jamaican, his unworthiness and inadequacy. But he cannot live with a self picture as unlovely as this. So he seeks reassurance by conceiving of himself as much better than he actually is. Hilgard[1] commenting on the operation of the mechanism in the Self Concept, points out that the need for self deception arises because of a more fundamental need to maintain or restore self esteem.

Perhaps, however, the explanation lies simply in the differing experiences of the respective groups. The Jamaican samples were drawn from a level of the population which, by and large, did not have many opportunities for getting ahead and achieving success. But these particular ones had achieved greater success than their peers. Thus, they saw themselves as more intelligent, of better personality, able to achieve more, etc. On the other hand, their English counterparts, were less successful than their peers from grammar schools who had gone on to university or entered other more highly esteemed professions. It should be noted, too, that though the Jamaican samples had a higher Ideal Self, they were, in fact, more Self Acceptant.

TABLE 4

SEX DIFFERENCES IN THE SELF CONCEPT.

	Men			Women			C.R.	Sig.
	N	M	σ	N	M	σ		
COGNIZED SELF								
Jamaican—Regular	43	172·58	10·96	57	176·50	12·80	1·64	N.S.
English—Regular	49	143·26	18·70	58	137·30	19·30	1·49	N.S.
IDEAL SELF								
Jamaican—Regular	40	181·75	14·62	49	195·26	9·45	5·03	·001
English—Regular	48	172·63	17·80	58	172·87	12·90	·078	N.S.
SELF ACCEPTANCE								
Jamaican—Regular	38	10·16	10·10	46	17·36	10·91	3·13	·01
English—Regular	48	29·92	14·50	58	35·02	14·48	1·81	N.S.

It had been expected that there would be no difference in the scores of men and women on the Self Concept scales at either levels of the Self or in Self Acceptance in either sub-culture. It was found that this was true, for the English samples, but not for the Jamaican samples. For the Jamaican samples, though the difference in the total scores of men and women did not reach significance, women cognized themselves significantly higher than men on three of the five areas making up this total score. On the Ideal Self women were significantly higher than men, and they were less Self Acceptant.

It is possible that the explanation of the difference in Self Concept of Jamaican men and women lies in the fact that the Jamaican society is somewhat matriarchal in structure, though in certain respects it is patriarchal of the Victorian flavour.* Women very often have to assume the role of both mother and father in the family. They are very often the mainstay and general provider. It is not surprising, therefore, that they should conceive of themselves in higher terms than their male counterparts. But perhaps, they are not satisfied with playing this role. This might account for their being less Self Acceptant than men. No similar pressures appear to be operating in the English population.

TABLE 5

AGE AND THE SELF CONCEPT.

	Regular			Mature				
	N	M	σ	N	M	σ	C.R.	Sig.
COGNIZED SELF								
Jamaican Female	57	176·50	12·80	88	166·32	17·00	4·17	·001
English Female	58	137·30	19·30	11	143·82	20·80	1·02	N.S.
IDEAL SELF								
Jamaican Female	49	195·26	9·45	84	190·61	10·70	2·60	·01
English Female	58	172·87	12·90	11	172·45	15·60	·10	N.S.
SELF ACCEPTANCE								
Jamaican Female	46	17·36	10·91	76	24·25	13·70	3·08	·01
English Female	58	35·02	14·48	11	27·91	13·50	1·51	N.S.

It was believed that by the time people have reached adult status, the Self Concept would have become relatively stable and would consequently change very little in succeeding years. It was, therefore, hypothesised that there would be no significant differences in the Self Concepts and Self Acceptance of 'regular' and 'mature' training college students in either culture. It was found that this was true of the English samples but not of the Jamaican samples. There are some interesting differences in the direction of change in the different areas however.†

It seems reasonable to conclude that by early adulthood the personality structure of students in the English sub-culture has achieved greater stability than the personality structure of students in the Jamaican sub-culture. Consequently, while the Self Concepts of English students change little with succeeding years, Jamaican students are still changing and modifying their concepts of themselves. Perhaps they are less mature emotionally than their English counterparts.

---

* See Kerr, M.[3]

†For full exposition, see Phillips, A. S., *loc. cit.*

VI.—REFERENCES.

[1] HILGARD, E. R. (1949). Human motives and the concept of self. *Am. Psychol.*, 4, pp. 374-384.

[2] JERSILD, A. T. (1952). *In Search of Self.* New York: Bureau of Publications, Teachers' College, Columbia Univ.

[3] KERR, M. (1952). *Personality and Conflict in Jamaica.* Liverpool University Press.

[4] LECKY, P. (1945). *Self Consistency: A Theory of Personality.* New York: Island Press.

[5] STAINES, J. W. (1958). The self picture as a factor in the classroom. *Brit. J. Educ. Psychol.*, 28, pp. 97-111.

(*Manuscript received 12th November, 1962.*)

VALUE OF THE ORAL RESPONSE IN BEGINNING READING: AN EXPERIMENTAL STUDY USING PROGRAMMED INSTRUCTION*

By JOHN D. McNEIL and EVAN R. KEISLAR
(*University of California, Los Angeles.*)

SUMMARY. The question of the effect of vocalization in reading has not received a clear answer. In the present study, the procedures of programed instruction were used in testing the hypothesis that the saying of words and sentences out loud by beginners in reading will result in greater ability to recognize and understand written words and sentences. An auto-instructional program which taught forty words was administered to 182 non-readers in the kindergarten during a seventeen-day period. These children had been paired and assigned to oral and non-oral groups where they received identical programed instruction with the exception that those in the oral group were directed to 'say the words' instead of to 'look at the words.' An examination which called for silent reading, not oral reading, was given at the end of the last lesson. The results supported the hypothesis that the oral response is valuable in learning to read.

I.—INTRODUCTION.

How important is it for beginners to say words out loud in learning to recognize and understand written words and sentences? In order to collect evidence regarding this perennial and fundamental question, procedures of programed instruction were used since they offer a high degree of reproducibility and control.

The psychological and educational literature has not given consistent answers to the value of oral responding in learning to read. Recognition and recall of objects have been found to be greater if the names of the objects are vocalized by the learner [2, 3, 6]. Jeffrey reported research which suggested that self-stimulation, the responding to one's own responses, helps in the overcoming of difficulty in learning to apply distinctive names.[5] One's own saying of a word might serve as a stimulus as well as a response. It is a stimulus in that the individual can hear and feel himself talk. "The feel of saying the word is part and parcel of the same pattern of speech stimulation."[4]

On the other hand, the observation that many poor readers verbalize has led some to the opinion that an oral response in learning to read is detrimental to reading achievement. Acting on this assumption, educators like McDade[7] and Buswell[1] sought favorable results from a method of beginning reading which required no oral responding and in which children were told not even to say the words to themselves. Learners were only to match words with objects, pictures, and actions.

The specific hypothesis of the present study was that kindergarten children taught by programed materials which required the saying of words and sentences out loud would achieve greater skills in reading than children taught by the same programed materials with no instruction to speak out loud.

---

* This research project was supported by a grant from the United States Office of Education, Department of Health, Education, and Welfare (Grant No. 1413).

II.—DESIGN OF THE EXPERIMENT.

The general plan of the study called for presenting an auto-instructional program, consisting of seventeen daily lessons, to a number of kindergarten children in small groups, half of whom were required to make oral responses and half of whom did not. At the end of the last lesson, the extent of learning was measured by a post-test. The presentation of the program to the small groups constituted replications of the experiment, providing evidence of reliability of the treatment and its effects.

Subjects.

The children used in the study had an average chronological age of 5 years and were the total kindergarten populations at three selected public elementary schools. One school represented an upper-middle class native-born white community; one school was located in a lower-middle class community, predominately of Japanese and Mexican backgrounds; and a third was located in a middle class community typical of those in most Southern California cities.

Before commencing the experiment, each child was given an intelligence test as well as a reading readiness test which measured his knowledge of letters, words to be taught in the program, and ability to recognize likeness in letter configurations. Children from the same classroom were paired on the basis of scores obtained from these tests and also upon their record of absenteeism. One member of each pair, randomly selected, was assigned to an oral group, the other member to a non-oral group. From these pairs thirteen oral and thirteen non-oral groups consisting of ten or less were constructed.

There were 231 children in the original sample. The subjects who were dropped from the study were: forty-three children absent more than three consecutive days, four children unable to speak English and two children identified by their teachers as hyperactive (emotionally disturbed). The remaining 182 children, therefore, made up the subjects for the study. These children had an average intelligence quotient of 107 with a range from 72–138 as measured by the Kuhlmann-Anderson Test of Intelligence.

III.—METHOD.

Programed Materials.

The instructional program consisted of seventeen daily lessons which took slightly over three weeks to administer. Each lesson was made up of approximately thirty-five questions or frames. Two sets of the programed materials, identical in all respects but one, were prepared. The non-oral set carried a taped commentary which directed the learner to "look at the word(s)"; the other set directed the learner to "say the word(s)." In order to assure that children in both groups were attending to the reading materials, all subjects were required to respond to problems throughout the program; for each problem, children were to select one of three choices offered by pressing an appropriate button. The oral group, therefore, made both oral and button-pressing responses; the non-oral group made only the button-pressing response.

Forty words were taught during the program both as independent words and as sentences. These words represented terms commonly found in the oral vocabulary of young children[8] and included both function words (i.e., and, with) and those with a direct referent (i.e., mother, girl). The printed words taught in this study were among those used in everyday speech by children. The study was not concerned with extending the child's vocabulary.

Examples of the 700 frames in the program are shown below :

	Purpose.	*Frame.*	*Commentary.*
(1)	To introduce a new word.	Children children woman dog	" Children. There are children in your room at school. The word at the top says children. (Say the word.) (Look at the word.) Which word below also says children ? "
(2)	To facilitate reading from left to right and to emphasize relations of a printed word to a pattern of language structure.	children run ...... fast girl boy	" (Say) (Look at) the words at the top. Which word below sounds right with the top words ? "
(3)	To maintain motivation as well as to test for comprehension.	the boy and girl can run picture picture picture A B C	" (Say) (Look at) the words at the top and find the picture that goes with the words."
(4)	To help the learner identify sentences which are appropriate answers to the questions he hears.	Billy Billy we go can eat runs fast	" What does Billy do when he is in a hurry ? "
(5)	To help learners rely on more than one cue in identifying words. (e.g., length instead of initial letter or configuration alone).	children cold chair	" Which word says children ? "
(6)	To give practice in recognizing and comprehending function words.	children ...... run we fast can	" Find the missing word."

The program aimed to develop four genaralizable skills. Beginning with the first lesson, efforts were made to teach children (1) to respond appropriately to new combinations of printed words, arriving at a complete thought or meaning, (2) to identify the particular word in a sentence which provides an answer to a posed question, (3) to respond to a variety of cues in discriminating among words, and (4) to read from left to right. To accomplish these four objectives, the following principles of program preparation were adopted :

(*a*) In order to get the learner to read from left to right, new words were introduced on the right of previously learned words (e.g., Frame (2) above).

(*b*) Pictures were rarely used in the program. The responses to printed words were prompted orally by the taped commentary. Occasionally, for motivational purposes, children would be asked to look at or say the words at the top and then find the picture from among three that " went with the word " (e.g., Frame (3)).

(*c*) The sequence of frames was built around reiteration of the words and the generalizable skills of word and sentence recognition. There was no narrative or thematic structure. Words and sentences were reviewed at unexpected times.

The overall difficulty level of the program (percentage of errors on first trial) was 33 per cent. for the oral group and 38 per cent. for the non-oral group. Performance by individuals reflected a range in difficulty from 70 per cent. errors to 2 per cent. The average difficulty of the daily lessons ranged from 23 per cent. to 43 per cent. errors (oral group) and 23 per cent. to 51 per cent. (non-oral group).

The Apparatus.

The apparatus provided for instruction of ten or fewer children at a time. Each child sat in an individual laboratory-type cubicle equipped with headphones through which he heard the taped commentary. The cubicle also contained a microphone as well as response and confirmation panels. The response panel consisted of three buttons by which the child indicated his selection in answering the multiple-choice problems in reading. Confirmation of the correct answer was by lights; a green light was shown when the learner made the correct response, and a red one when he was in error.

A single 6-ft. × 8-ft. screen at the front of the room was visible from each cubicle. On this screen, learners saw each daily sequence presented by film-strip, while simultaneously hearing the accompanying taped commentary. The program advanced automatically at a paced schedule with 3-seconds allowed for responding to each frame. As each individual in the oral group made vocal responses into his microphone, he heard his own voice. Frequent monitoring of the individual oral responses through the microphone circuit gave evidence that subjects were responding orally as directed. When the non-oral group used the equipment, the microphones were deactivated. A record of each individual's multiple-choice performance for each frame was obtained by means of a punched paper tape.

The apparatus, therefore, combined advantages of both group teaching (i.e.. economy, control of word exposure time and pacing) and tutorial instruction (i.e., personal communication without peer group distraction and provision for individual response and confirmation).

The Testing Procedure.

The criterion test provided evidence that the learner had acquired knowledge of the words and reading skills previously mentioned. This post-test consisted of fifty-one multiple-choice items, each with three or four alternatives, similar in format to the frames in the program except that the items appeared in individual booklets, one item per page. Instructions for each item were given orally. No item in the post-test was identical to any frame in the program; the context and the arrangement of alternatives had not been seen previously by the pupil. Twenty-four items required that the learner recognize a given word. The remaining items required the child to select the individual word or sentence which supplied the answer to an oral question, or an incomplete sentence, or which best described a picture.

Examples of the test items are as follows:

Objective.	*Instruction.*	*Test Item.*
He selects an individual word when it appears with those of (*a*) like and (*b*) unlike shape.	" Find the word that says dog."	(*a*) boy, dog, toy. (*b*) children, tree, dog.
He selects the word that will complete a sentence never seen before.	" Find the missing word."	a dog can play ...... a ball red, with, no, little

The test was administered in group situations, twelve children at a time, oral and non-oral together.

IV.—RESULTS AND CONCLUSIONS.

(1) *Overall Effect of Oral Response.*

The mean score for the oral group on the criterion test was 30·3 (SD=10·0) while the non-oral mean was 23·8 (SD=7·9). This superiority on the part of the oral responders was significant when based on the average difference between matched pairs ($t=5{\cdot}64$, $p<{\cdot}01$). Since the mean score for forty-seven uninstructed kindergarten children was only 12·7 (SD=3·3), both the oral and non-oral forms of the program appear to be effective in teaching.

The superiority of the oral response group is illustrated in Figure 1 where the results are presented for the three groups in terms of percentage of total possible scores. Since there was no correction for guessing, the dotted line indicates a chance score which would have been obtained by blind guessing. Using this line as a base, it may be noted that the effect of the oral response is considerable.

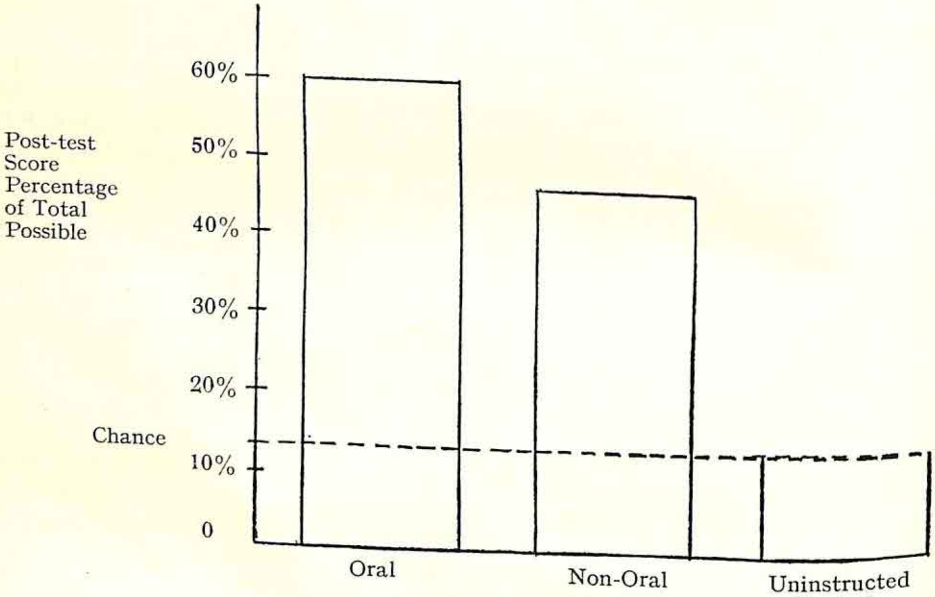


FIGURE 1

COMPARATIVE CONTRIBUTION OF ORAL AND NON-ORAL RESPONSE MODE IN READING ACHIEVEMENT.

Since the experiment was carried out with thirteen pairs of matched groups, we may also view the data non-parametrically. In each of the thirteen pairs of groups, the mean of the oral group was higher than the mean of the corresponding non-oral group. On the basis of a sign test, this difference is statistically significant ($p<{\cdot}01$).

(2) *Value of Oral Responding at Different Levels of Ability.*

Although it was speculated that the oral response would be especially important for children of lower intelligence, the interaction effects were not large enough to be statistically significant. However, the direction of the results suggest the desirability of further inquiry regarding the special advantages of oral responding by those of lower intelligence. Table 1 below shows the mean scores by three levels of ability for both oral and non-oral responders.

TABLE 1

Post-test Results for Oral and Non-Oral Subjects.

	Oral Written Post-test	Non-Oral Written Post-test
Low I.Q. (N=60)	M=25·2	M=17·0
Middle I Q. (N=62)	M=31·8	M=24·1
High I.Q. (N=60)	M=33·7	M=30·3
Total N=182	M=30·3	M=23·8

(3) *Value of Oral Response in Motivating Children to Want to Read.*

In order to throw light upon the attitudinal effect of the program and the oral response, an interview was held with each child. Children from both oral and non-oral groups were individually asked whether they would prefer to continue with the program when given each of the following as an alternative: (1) listen to records; (2) listen to a story; (3) paint; or (4) play in the sandbox. Over 72 per cent. of the children in both groups preferred the reading activity to one or more of these other activities; 10 per cent. chose the reading activity in preference to all four of the other alternatives. Parenthetically, we can say that the introduction of a formal reading program was received with as much enthusiasm as currently popular activities in the kindergarten.

Although achievement in learning was not found to be associated with preference for continuing the reading activity, the data suggested that oral responding might have motivational properties for those of lower intellectual ability (I.Q. below 100). Preferential scores for oral and non-oral groups in middle and high I.Q. levels averaged 1·5, but preferential scores for the low I.Q. level were 1·6 and ·97 for the oral and non-oral groups, respectively.

(4) *Conclusion.*

The results show that oral responding by beginning readers facilitates the recognition and comprehension of printed words and sentences. Children who responded orally during the auto-instructional program were superior on a test which called for the application of a variety of reading skills in contexts different from those taught in the program itself. The oral response was of significant value even though the post-test called for silent reading, not oral responding, and none of the items of the post-test appeared in the training program.

As indicated earlier, one way to explain the mechanism involved in the greater achievement with overt vocalization is that the distinct stimulus of saying words aloud is required for mediation of the obtained effects.

The experiment gives no support for the belief that oral responding in learning to read has unfavorable effects. On the contrary, we must conclude that, although vocalization and poor reading on the part of the learner are often associated, the act of vocalization during instruction is an aid to the beginning reader in what for him is already a difficult reading situation. Without such vocalization, his performance would be worse.

V.—References.

[1] Buswell, G. T. (1945). Non-Oral reading: a study of its use in the Chicago Public schools. Chicago, Illinois: *Supplementary Educational Monograph, No.* 60.

[2] Cantor, G. N. (1955). Effects of three types of pre-training on discrimination learning in pre-school children. *J. Educ. Psychol.*, 49, 339-42.

[3] Dietze, D. (1955). The facilitating effect of words on discrimination and generalization. *J. Educ. Psychol.*, **50**, 255-60.

[4] Horn, Ernest, and Curtis, James F. (1949). Improvement of oral reading. *N.S.S.E. Forty-eighth Year Book, Part II*, p. 254.

[5] Jeffrey, Wendell (1961). Perceptual learning in children. *Research Roundup on Children and Youth*, University of California Press.

[6] Kurtz, K. H., and Hovland, C. I. (1953). The effect of verbalization during observation of stimulus objects upon accuracy of recognition and recall. *J. Educ. Psychol.*, **45**, 1957-64.

[7] McDade, J. E. (1937). A hypothesis for non-oral reading: argument, experiment, and results. *J. Educ. Res.*, 30, 489-503.

[8] Stone, C. R. (1956). Measuring difficulty of primary reading materials: a constructive criticism of Spache's measure. *Elem. Sch. J.*, LVII, 1, 36-41.

(*Manuscript received* 19*th November*, 1962.)

THE INFLUENCE OF EARLY PUBERTY ON TEST PERFORMANCE AT AGE ELEVEN

BY

J. D. NISBET

(*Department of Education, University of Aberdeen*)

AND

R. ILLESLEY

(*Obstetric Medicine Research Unit, University of Aberdeen*)

SUMMARY. It has been suggested that the earlier onset of puberty among girls may now be affecting performance in tests at the age of 11 plus. This question was examined by the analysis of test scores at ages 7, 9, 11 and 13 from a year-group of girls numbering 1,385, for whom age at first menstruation and father's occupation were known. The very early maturers proved to have a slightly higher average test score than the others, but they were superior at all ages and not only at 11. While the results at age 13 do not rule out the possibility of a later pubertal spurt, it can be concluded that early onset of puberty does not confer any advantage in selection tests at age 11.

I.—INTRODUCTION.

THE evidence for the secular trend towards early maturity has been summarized recently by Tanner (1961): for example, he reports that among girls in Norway, in 1840, the average age at menarche (or first menstruation) was 17½ years, whereas to-day it is about 13 years. The trend appears to have been maintained steadily over the whole of the last hundred years, and it is still continuing. (In earlier times, menarche seems to have occurred around the age of 14. Shakespeare's Juliet was 13: he devotes ten lines of Act I, Scene III, to emphasize that she is just "a fortnight and odd days" short of her fourteenth birthday.) The factors influencing the age of puberty remain uncertain; inadequate records and differences in mortality rates make research difficult.

This trend has aroused concern among those who are responsible for operating procedures of selection for secondary education. Girls reach puberty earlier than boys; and if there is a spurt in mental growth similar to the acceleration in physical growth at puberty, then as puberty occurs earlier, the spurt will be beginning to affect the performance of early maturing girls in tests at the age of 11 plus. Since the advantage of an early growth spurt is only temporary, this would reduce the validity of the selection procedure. Tanner (1962) has recently reported that early maturing girls are, indeed, scoring higher on average in selection tests than the late maturers. However, this finding does not allow for the possibility that the early maturers may have a consistent superiority at all ages over the late maturers.

Other indirect evidence would appear to be consistent with the hypothesis that scores at age 11 are beginning to be affected by the earlier onset of puberty, though here too other explanations of the findings are possible. The Scottish Mental Survey (1949), in which a test given to all the 11-year-old children in Scotland in 1932, was given again to the 11-year-old age-group in 1947, showed that the average score in 1947 was higher than in 1932. In 1932, the boys were slightly better on average than the girls. By 1947, the girls had not only caught up, but had passed the boys, and scored on average two points higher than the boys. Emmett (1950) in a similar comparison of age-groups in ten English areas over the period 1936-1947, also found that the average score of girls had increased relative to that of boys, though there was some indication that the girls' superiority had begun to fall away by the age of 13. Boyne and Clark (1959) found a similar increased superiority of girls over boys among Aberdeen school children over the period 1947-1957.

These findings would appear to be in accord with the hypothesis that a spurt in mental growth accompanying puberty is affecting performance at age 11. However, the data can be explained without assuming an adolescent spurt in mental growth. Differences between mean test scores of boys and girls may arise from a variety of causes. For example, girls are generally better than boys in verbal problems and boys are generally better in non-verbal and numerical problems. There is a tendency in many tests to use non-verbal material for the simpler items and verbal material for the more difficult items. Thus, if there is a slow general improvement in over-all performance (whether due to coaching or any other cause), this would also result in a trend towards girls' superiority as the whole age-group moved up into a type of item which tended to favour the girls.

For an adequate test of the hypothesis, a longitudinal study is necessary. An age-group must be tested at intervals from age 7 (or from any age at which it can be safely assumed that puberty has not yet affected test performance), up to at least age 13 (by which age it can be assumed that the influence of puberty, if it exists, will have begun to show an effect). Comparison of the average scores of the early maturers and the remainder provides a test of the hypothesis that there is a spurt in mental growth accompanying puberty, and will show whether earlier onset of puberty is affecting performance in tests at age 11.

II.—SUBJECTS, TESTS AND OTHER DATA.

The population studied was defined as those girls who were in the second year of secondary schools in Aberdeen—that is, around the age of 13 or over—and who were present on a day in December, 1960, on which one of our tests was given. For this year-group, the education authority records were searched to identify the earlier scores of these girls in the tests which had been given to the whole age-group at ages 7, 9 and 11, as part of the regular testing programme in the City of Aberdeen. On these earlier occasions, absentees were followed up to ensure that a test score was available for every pupil at each age; for lack of resources it was not practicable to follow up the absentees at age 13. Although all secondary schools, both local authority and independent schools, were included in the testing at age 13, earlier test scores were not available for most of the girls from two private schools attended by some 6 per cent. of the age-group of girls, as only a few of the girls at these two schools had taken the earlier tests. Migrants into the city also missed the earlier tests. In the end, complete test records from age 7 were traced for 85 per cent. of those tested at age 13. The sample is thus slightly biased towards the lower occupational groups.

The tests used were: at age 7, Moray House Picture Test; at age 9, Schonell's Essential Intelligence Test; at age 11, Moray House Verbal Reasoning Tests 57 and 58; and at age 13, Moray House Advanced Verbal Reasoning Test 10. Not all the sample sat the earlier tests at the same time: thus, at age 11, the large majority sat Moray House Tests 57 and 58, but a few sat the tests applied the previous year, Tests 55 and 56. This difference had to be ignored: age-allowances provide for differences in age at testing, and the Moray House Tests are intended to be, and generally are, comparable with one another. All the test scores were re-standardized on this sample at each age, to equate the means and standard deviations and to produce normal distributions in each test, thus eliminating differences in standardization between tests. Means of 100 and standard deviations of 15 were imposed in a conversion based on scores grouped in intervals of five points: interpolations within intervals resulted in means which worked out slightly below 100 and standard deviations below 15. The means and standard deviations of the original scores and of the scores after conversion are set out in Table 1.

TABLE 1

MEANS AND STANDARD DEVIATIONS BEFORE AND AFTER RE-STANDARDIZATION.

Test at Age—	Original Scores		Revised Scores	
	Mean	s.d.	Mean	s.d.
7	106·85	15·33	99·77	14·66
9	109·70	15·53	99·66	14·81
11 (1)	104·95	13·68	99·82	14·82
11 (2)	106·41	13·44	99·66	14·69
13	107·90	13·93	99·49	14·89

The next step was to find out age at menarche for each girl. The investigation was restricted to girls so that the age at first menstruation could be used as an indication of whether puberty came early or late. Since the sample was aged about 13 or over, menarche was a fairly recent event for many of them and it was, therefore, not necessary to rely on distant and possibly faulty memory. A considerable proportion had not yet reached menarche, but it was still possible to separate the early developers and the very early developers from the remainder. Existing medical records did not include age at menarche; but the School Health Service in the city undertook to interview individually all the girls in the sample to ask for this information. In three schools, the head preferred that the information should be got by writing to the mothers of the girls. This was done and practically all replied. Difficulties at this stage of the inquiry had been anticipated and particular attention was given to the method of approach. No one was scandalized, no protests were received, and information on age at menarche was obtained for 98·8 per cent. of those tested at age 13.

Information on social class was based on the answers to carefully designed questions on father's occupation. On this basis, pupils were graded into six groups on a modified version of the Registrar-General's Scale: I and II, professional and business; IIIa non-manual skilled workers; III skilled manual workers; IV semi-skilled manual workers; and V unskilled workers. When a pupil's description of the father's occupation was too vague to be classified, the pupil was asked, through the headmaster, for further information. The answer "Don't know (father's occupation)" was permitted, to allow for illegitimate children and other special cases. Only 2·3 per cent. of the total group could not be given a grading in one of the six categories and it was possible to establish from the detailed records of the Aberdeen Maternity Hospital that a considerable proportion of these were, in fact, illegitimate children.

III.—ANALYSIS OF RESULTS.

Complete test records—that is, scores on tests at ages 7, 9 and 13, and on two tests at age 11—were available for 1,402 girls. Information on menarche was lacking for seventeen of these, leaving a total of 1,385, from whom the following results are derived. The age range at the time of interview included six pupils who were under 13 years, and six over 14 years 6 months, but 91 per cent. of ages at interview were between 13 years 3 months and 14 years 2 months. The median age at menarche in the sample was 13 years 2 months. Menstruation began before the age of 11 in twenty-four cases, under 2 per cent.; before age 12 in 165, or 12 per cent.; and before age 13 in 40 per cent. of the group. Age at menarche was classified in three categories:

(i) The very early maturers, with menarche before the age of 12 years 3 months.

(ii) The early maturers, with menarche between the ages of 12 years 3 months and 12 years 11 months.

(iii) The remainder, among whom menarche occurred at or after the age of 13 years, or who, being already aged 13 years, had not yet reached menarche. (There were two girls not yet at menarche who were aged between 12 years 9 months and 12 years 11 months, and these were included in this third group).

This third group, therefore, included both the median group and the late maturers. It was not practicable to distinguish these. There was, in the sample, a group of 358 girls among whom menarche occurred after the age of 13 years 9 months. The mean scores of this group proved to be about five points of I.Q. below the average, but this was a spurious result arising from the nature of the sample. Where a sample is drawn from an educational year-group, there is a tendency for the brighter pupils to be younger than average and the duller pupils to be older. Younger pupils who were to prove late developers had not reached menarche at the time of interview. An attempt to establish a category of definite late developers from the available data would have missed these younger pupils and resulted in a group with a bias towards older age and lower ability. All pupils who were neither early nor very early maturers were, therefore, grouped together.

The same objection does not apply to the 'early' and 'very early maturers,' with whom menarche occurred before age 13 years or before 12 years 3 months, respectively, as these groups could be identified with certainty. It is thus possible to derive results which would test the hypothesis that the early maturers would show a spurt in mental growth around the onset of puberty, but not to obtain information on the effect with late maturers. It is, however, the early maturers who are concerned in the argument about the influence of puberty on performance in selection tests at the age of 11. The hypothesis under test is that, if there is a spurt in mental development with the onset of puberty which is beginning to affect performance in testing at age 11, this would appear as a sudden improvement of the very early maturers at the age of 11—an improvement, that is, relative to any possible superiority that they may previously have shown over the early maturers and the remainder at the age of 7 and 9 before any effect of puberty has begun to operate. By age 13, on this hypothesis, the lead of the very early maturers over the earlies would have been reduced, as the earlies, and perhaps some of the remainder, came to benefit likewise from this pubertal spurt.

The mean scores of the very early and early maturers and the remainder, which are set out in Table 2, do not show this pattern of results. The very early maturers are certainly superior at the age of 11, but this is not a sudden emergence of superiority with the onset of puberty. They are almost the same degree ahead at ages 7 and 9. In fact, the pattern of mean scores between ages 7 and 11 is, if anything, the reverse of what would be expected on the hypothesis of a pubertal spurt among very early maturers at age 11. Differences between the means of the menarche categories at any one age are generally significant (the 'very early' at all ages over both 'early' and 'remainder'; the 'early' only at ages 7 and 13 over 'remainder').

TABLE 2

MEAN TEST SCORES OF MENARCHE GROUPS.

Mean Test Score at Age—	Menarche		
	Very Early	Early	Remainder
7	103·5	100·7	98·3
9	102·8	99·4	98·8
11 (1)	102·9	99·9	98·9
11 (2)	102·6	99·8	98·7
13	103·3	100·2	98·0
N	246	300	839
Percentage entering Senior Secondary Schools	28	23	21

Beyond the age of 11, with the results at age 13, there is some suggestion that the early maturers are pushing ahead of the remainder. The differences between means at various ages for any one menarche category are not significant until we reach age 13: at this age, the drop in mean score of the 'remainder' category is significant. An alternative expression of this point is that the distance between the means of the 'very early' and 'remainder' categories increases significantly between 11 and 13. This may mark the beginning of a pubertal spurt among the very early maturers, though without a further testing at age 15 (now impracticable as many of the pupils have already left school) the data are inadequate to establish such a finding. However, it can be concluded from the present data that the earlier onset of puberty among girls is not seriously affecting test performance at age 11.

It should, perhaps, be emphasized that within each menarche group there is a wide range of test scores at each age, the very early, early and remainder categories all including pupils whose scores range from the highest to the lowest.

The relation of social class (assessed by father's occupation) to age at menarche and test score at various ages can be examined from the data. There is, of course, an association between social class and test score: the average of girls from Social Class I was between 114 and 117 on all the tests, and the average score of those from Social Class V was between 93 and 95. Again, there was a wide range of scores in each category. The superiority of the very early maturers cannot, however, be explained as merely the product of class differences. Even within any one social class, the very early maturers are consistently above the average for that class. The mean scores for each menarche category, separating the social class groups, are set out in full in Table 3.

TABLE 3

MEAN TEST SCORE AND MENARCHE, BY SOCIAL CLASS.

Social Class	Menarche Category	Mean Score at Age—					N
		7	9	11 (1)	11 (2)	13	
I/II	VE	115	116	114	115	116	(35)
	E	107	107	109	109	108	(47)
	R	108	107	108	107	108	(123)
IIIa	VE	107	106	105	105	106	(46)
	E	105	102	104	105	106	(37)
	R	97	101	103	103	101	(116)
III	VE	101	99	100	100	100	(111)
	E	101	99	99	99	99	(124)
	R	98	98	98	97	97	(372)
IV	VE	103	102	100	99	100	(22)
	E	94	96	96	96	97	(44)
	R	97	95	95	96	96	(98)
V	VE	97	97	99	98	98	(28)
	E	97	93	94	93	93	(43)
	R	93	94	92	92	92	(107)
Social class not stated : VE 4 ; E 5 ; R 23							(32)
					Total	....	1385

Sub-division by social class reduces the numbers in individual categories of the table, resulting in slightly larger chance fluctuations in the means at various ages. However, breaking down the data by social class produces no evidence to affect the conclusion derived from Table 2. Within each social class, the very early maturers generally have higher mean scores than the rest, but this appears at all ages and not merely with the onset of puberty.

Some association between social class and age at menarche might reasonably have been expected. When the investigation was being planned, the physiologically-minded among our advisers were inclined to predict that the early maturers would tend to come from the upper social groups, being taller, heavier and quicker to mature because of better nutrition. Some teachers, on the other hand, pointed out that girls from the poorer urban areas give the impression of earlier maturity in their appearance, dress, behaviour and interest in boys, and, in fact, they tend to marry earlier. Table 4, which presents the numbers and proportion in each menarche category for each social class, does not support either of these views : there is no marked association between age at menarche and social class in this sample. The proportion of very early and early maturers is nearly the same among daughters of professional and business men as among the daughters of unskilled labourers.

The data for Table 4 were drawn from all those for whom information on both social class and age at menarche was available, that is, 1,385 girls on whom the previous tables are based, less thirty-two for whom a social class grading was not available, plus 211 girls for whom test records were incomplete (mostly from private schools and social classes I and II). A parallel analysis of the 1,385 cases gives similar results.

TABLE 4

DISTRIBUTION OF MENARCHE BY SOCIAL CLASS.

Menarche Category	Social Class					Totals
	I/II	IIIa	III	IV	V	
Very early	50 (16)	49 (21)	119 (18)	24 (13)	31 (16)	273 (17)
Early	67 (22)	44 (19)	131 (20)	48 (26)	46 (24)	336 (21)
Remainder..	189 (62)	135 (59)	404 (62)	110 (60)	117 (60)	955 (61)
Totals......	306 (100)	228 (100)	654 (100)	182 (100)	194 (100)	1564 (100)

(Percentages in brackets.)

Chi-square on this table is 8·6, with 8 degrees of freedom, and is not significant (P about 0·4).

The inter-correlations between the test scores at the various ages were as follows:

Between score at age 7 and score at age 9 : +0·69
age 11 (1) : +0·72
age 11 (2) : +0·75
age 13 : +0·74
Between score at age 9 and score at age 11 (1) : +0·82
age 11 (2) : +0·82
age 13 : +0·82
Between score at age 11 (1) and score at age 11 (2) : +0·93
age 13 : +0·90
Between score at age 11 (2) and score at age 13 : +0·91

IV.—CONCLUSIONS.

The results from this analysis of the test scores of 1,385 girls at ages 7, 9, 11, and 13, show that the anxiety that earlier onset of puberty is affecting performance in selection tests at age 11, is unfounded. On other aspects of the effect of puberty on mental development, the data do not permit any definite conclusions to be reached. The results at age 13 can be interpreted as indicating the beginning of a possible later spurt. It is also possible (some may say probable) that the onset of puberty affects aspects of mental development which are not measured by the tests used. Moreover, the investigation was concerned principally with a comparison of the early and very early maturers with the remainder, and the data were not appropriate for a study of the late maturers. Whatever the causes of the earlier onset of puberty, the evidence so far does not show any relationship between early onset of puberty and social class. Other aspects of this investigation, including family size, migration, type of secondary school attended, area of residence and birth weight, will be discussed in a subsequent paper.

Acknowledgements.—This study was begun in collaboration with David Hewitt, of the Department of Social Medicine, Oxford University, who was prevented by absence abroad from participating in the analysis of results. It will be obvious that the investigation involved the collaboration of a large team of workers, who assisted in the planning and organisation, in the administering and marking of tests, in the collection of data, in maintaining the records from which data were drawn, in searching these records, in coding and sorting the data of the investigation proper, and in providing information and assistance generally. Acknowledgement is due to the Health Visitors in Aberdeen under Dr. I. A. C. MacQueen, Medical Officer of Health; Dr. D. Younie, of the Aberdeen School Health Service; the headmasters, head-mistresses and teachers in Aberdeen secondary schools; Mr. W. Boyne of the Rowett Research Institute, Aberdeenshire; Mr. J. R. Clark, Director of Education, Aberdeen, and Mr. Johnson and Mr. Argo on his staff; Miss Ann Lamont and other Ed.B. students in the Department of Education, University of Aberdeen; and the clerical staff of the Obstetric Medicine Research Unit in the University of Aberdeen.

V.—References.

Boyne, A. W., and Clark, J. R. (1959). Secular change in the intelligence of 11-year-old Aberdeen school children. *Human Biology*, 31, 325-33.

Emmett, W. G. (1950). The trend of intelligence in certain districts of England. *Population Studies*, 3, 324-37.

Tanner, J. M. (1961). *Education and Physical Growth*. London: University of London Press. Chapter 7.

Tanner, J. M. (1962). As reported in *Times Ed. Supp.*, Feb. 23, p. 355.

The Trend of Scottish Intelligence (1949). Publications of the Scottish Council for Research in Education, 30. London: Univ. of London Press.

(*Manuscript received 20th November, 1962.*)

THE MEASUREMENT OF LEARNING ABILITY

By G. W. S. MACKAY and P. E. VERNON
(*University of London, Institute of Education*)

SUMMARY.—Intelligence tests are often assumed to predict a child's ability to learn, although they are seldom based on situations that directly involve any learning. A battery of nine group tests was devised for junior school children, each based on the learning of an unfamiliar task, and yielding six scores for successive stages of practice. These were applied to 8-9 and 10-11 year pupils, along with a number of more conventional tests, in order to find whether they added appreciably to the predictions of achievement derived from the previous year's work. As a measure of learning, residual gain was used, i.e., final scores holding initial scores constant.

These residual gain measures were found to be largely specific in the younger group, though somewhat more consistent and more predictive in the older group. However, the tests differed considerably in their validities, and further work might produce more useful ones. There are considerable difficulties in finding tasks relevant to school progress which will also cover a wide range of ability. A stronger case might be made for using final scores, after a period of instruction and practice on new tests, as predictors.

I.—INTRODUCTION.

THE main object of giving intelligence tests to children in school, or at an educational clinic, is to assess their capacity to make scholastic progress. This implies that the tests predict ability to learn; and, indeed, several writers have defined intelligence as learning ability (Buckingham 1921, Dearborn 1921). If, however, we regard learning as changes in performance following instruction and practice, then, with Woodrow (1946), we "cannot simply state that improvement with practice is identical with some hypothesized intelligence." Much the same point is made by Simrall (1947), who adds that there is little or no evidence that gains due to practice on any one test are associated with gains on other tests; i.e., there is no such thing as generalized learning ability. Intelligence is normally measured by tests which are given once only, and which provide no opportunity for practice or the acquisition of new mental skills. Hence, they bear very little relation to the amount of progress children are likely to make in their school work over a given period, say a school year. And yet they still correlate highly with future overall scholastic achievement. J. E. Anderson (1940) has pointed out the main explanation of this apparent paradox, namely that mental development is always cumulative. A child starts the year with a fund of information and mental skills, X, which are very similar to those involved in doing an intelligence test, I; during the year he acquires further skills, Y. Now the correlation of I with final achievement, X+Y, is still quite large because r_{IX} is large, although the correlation r_{IY} may be very small or even zero.

Surely, it should be possible, then, to devise tests which actually involve learning in a controlled situation, and which might add to the predictions of later performance already covered by intelligence tests. Such tests should be of considerable value to the psychologist and/or the teacher. Furthermore, they should provide even better opportunities than current intelligence tests for observing and assessing children's methods of tackling difficulties, their attention, perseverance, etc.—qualities which are highly relevant to future success or failure.

II.—PREVIOUS INVESTIGATIONS.

A considerable amount of work on the factors involved in various types of learning has been carried out, though little of it has much bearing on scholastic achievement. Anastasi (1930, 1932), Thurstone (1938) and others, have demonstrated the existence of rote memory factors in the learning of paired associates, words, digits, etc.; for the most part, these seem to be fairly specific to the type of material and the conditions of presentation and recall. There is no evidence that they correlate with school learning, except in so far as they involve *g* or verbal ability. Vernon (1961) claims, in his review of the field, that logical memory or the learning of meaningful materials depends wholly on these familiar factors, i.e., that no special faculty of meaningful learning can be distinguished.

Another approach is represented by the investigations of Melton (1947) and Fleishman and Hempel (1954, 1955) into psychomotor learning, where measures of proficiency at different stages of practice were factorized, and changes in factor patterns were obtained as the learners became more proficient. In the early stages of practice at complex co-ordination or discrimination reaction tasks, there were considerable correlations with spatial, mechanical and other cognitive factors, suggesting that learners are tackling the task by means of their already established abilities. But at later stages there is less dependence on these familiar factors, and other motor factors become more prominent, or a factor specific to the particular task emerges. This explains why it is so difficult to predict ultimate motor skills either by paper-and-pencil aptitude tests or from success during the early stages of learning. A similar investigation of learning a short musical theme was published by Burroughs and Morris (1962).

Allison (1954, 1956) and Mollenkopf (1957), as part of a programme for selecting U.S. Navy personnel, designed performance tests involving the learning of mechanical tasks, in which subjects were instructed in assembling various models, and tested after each period of instruction. They found that these learning scores predicted achievement at the end of mechanical training at least as well as the Navy's Basic Battery of aptitude tests, and that they were capable of adding appreciably to the predictive efficiency of this Battery. Although no similar investigation has been carried out with scholastic materials, these results seem to offer some promise that observation and measurement of a pupil's progress in a learning situation might likewise improve on the prediction of his future achievement. G. R. Ortar (1960) was faced with the educational selection and guidance of Israeli children from very varied cultural and linguistic backgrounds. Finding that scores on standard tests were seriously affected by such differences, she introduced a period of coaching and, on retesting, obtained much improved validities in predicting school achievement. She did not, however, make use of measures of the gain brought about by coaching.

The most quoted studies of actual learning gains are those of Woodrow (1938, 1939), who practised two groups of subjects repeatedly on a number of simple tasks—cancellation, arithmetic, anagrams, length estimation, etc. Here also the factor content of the initial and final scores differed considerably, and there was little evidence of any general factor among the gain scores, nor of overlapping between the amount of improvement and intelligence. Nevertheless, gain scores on different types of task, e.g., numerical, perceptual, did show some positive correlation with *N* and *P* factors. One reason for Woodrow's negative results may have been that the tasks employed were of a rather simple kind, and Tilton (1953) claims considerably greater consistency in an investigation of school subject learning. Here, standardized achievement tests in eight subjects were given to fourth grade pupils, and repeated twenty months later. There were marked positive correlations, averaging 0·30 (or 0·51 when corrected for attenuation) between the gain scores, indicating a prominent general factor in school progress. Gains also correlated positively with initial achievement; i.e., those who were more able in any subject tended to progress better in it than those who were initially weak.

More recently, two large-scale investigations have been reported from the Educational Testing Service, Princeton, N.J. Stake (1958) applied a dozen learning tasks, each involving several trials, to 240 seventh grade pupils. Both verbal and non-verbal, rote and meaningful, tasks were used, and an extensive battery of reference tests was added. Each learning task was scored for total errors or overall success, for slope of the learning curves, and for regularity of progress. The last set of measures yielded mostly negligible correlations, but the success measures correlated positively with intelligence and with scholastic achievement. On factorizing his complete battery, Stake concluded that there is no general learning ability, but a number of factors for different kinds of tests; also that there is little relation between slope or rate of learning and familiar psychological factors.

Allison's (1960) study was similar, but it included a greater variety of learning tasks (psychomotor, mechanical, rote learning and concept formation), which were given to 315 adult recruits. These were scored for average rate of progress, and for faster or slower progress in the second than in the first half of trials. The main factors among these measures included: (1) Conceptual learning; (2) Mechanical and motor learning; and (3) Rote learning. On comparing these with the factors established in a battery of thirty-nine reference tests, he observed, unlike Stake, that No. 1 was related to general verbal intelligence and number ability, No. 2 to a spatial-mechanical factor, and No. 3 to a rote memory factor.

It would seem then that learning progress is certainly not identical with intelligence, though there may be some positive overlapping when complex learning tasks are studied. There is little evidence, also, of a general learning ability, common to all types of task, though this too may be affected by the level of complexity of the tasks. Rather, learning measures tend to group under a number of rather ill-defined factors, according to the type of material and the conditions of learning and recall. Some of these group factors do seem to relate to ability factors, for example verbal learning correlating more highly with verbal ability, non-verbal learning with spatial ability, etc.

III.—THE PRESENT EXPERIMENT.

Our own small-scale experiment was conducted against this background with two groups of approximately thirty-six primary school children, aged 8–9 years and 10–11 years, being the complete unstreamed second and fourth years of a school in a good middle-class area. A study with these numbers can be regarded only as a first pilot exploration; but it may suggest ways of tackling a larger sample with more refined techniques if worthwhile results are indicated.

The aim was to devise a series of group learning tests which would approximate more closely to the normal school learning situation than those employed by Stake. Group tests were preferred, not only because the tests would inevitably be too lengthy for individual application, but also because children might well learn differently when individually instructed from the way they learn in class. Each test was completed in a school period; i.e., long-term learning and retention were not involved. All tests were given during January, 1961, by the first author who, as head master, had often conducted classes with these children. In addition, a series of conventional 'reference' tests of verbal, numerical and spatial abilities and scholastic attainments was given, partly with the aim of exploring the factorial content of the new learning tests, partly in order to find whether the learning tests added significantly to conventional tests in predicting achievement. The final criterion of achievement consisted of a combination of the class teacher's estimates of progress made during the remainder of the school year, and standardized English and arithmetic tests.

An inherent difficulty in devising learning tests is that the more closely they parallel school learning, the more they are likely to be affected by previous experience and achievement. Another problem is to ensure that they will cover a wide enough range of ability to be useful throughout the junior school. Ideally, a dull 8 or even 7-year-old should be able to make a start with them before he receives practice; while at the same time a bright 11-year-old should still be improving at the end of practice. A description of the tests will show how these difficulties were tackled. Details may be found in Mackay (1962).

(1) *Language Learning Test*. This test attempted to teach the children a new language that they could translate from or into. Esperanto was used because of its inherent logical structure, and because it seemed unlikely that any children would be familiar with it. The test was a printed one and was tackled without any time limit, although it became clear that after about 40 minutes, the children had reached their limit. The initial instructions were not standardised; they were designed to reduce anxiety by giving the impression that this was a normal lesson, not a formal test. But once the nature of the task had been pointed out, the children carried on by themselves from printed instructions.

The first exercise showed a few easy English words with their Esperanto equivalents. Those carrying the definite article were prefixed with La, those with the indefinite article merely had the equivalent without La. This exercise expected the children to grasp this rule, and they had to use it in translating the given words. The rule was then carried forward to the next piece of information, and so on until sentences could be translated using the rules governing the pronouns, verb tenses, plurals, agreement of adjectives with nouns, etc. Finally, a complete piece of translation could be, and was, attempted by the children. In order to give an initial, final, and total score, the first third, last third, and total number of answers were marked separately.

(2) *Number Substitution Learning*. This test was similar in structure to the Language Learning. A simple form of algebra unfamiliar to the children was used. Letters were given equivalent numbers, and after some straightforward interpolation, additional rules were supplied involving addition, subtraction, multiplication and division, which, although familiar to the children, did depend upon previous experience. It seemed impossible to deal with numbers in any complex way without assuming a knowledge of these processes. The rules, however, had to be applied in a way that was not familiar, and it was the interpretation and progressive learning of the rules that was primarily involved.

(3) *Verbal Instructions Learning*. This test was given six trials, as was every subsequent test. The children were provided with six printed sheets, each sheet containing eight rows, each row consisting of two numbers and two letters. The tester read the instructions for all eight rows: e.g., put a ring round number seven and cross out letter A. After hearing the complete list, the children answered as many items as they could on their first paper. The full instructions were read again, and a second attempt was made with a new sheet. This was done six times.

(4) *Verbal Information Learning*. The above test involved rather meaningless material. This test presented a kind of learning more similar to what children meet at school. Twenty-five pieces of information were read out; these were simple, but unlikely to be known by any of them: e.g., Bach died in 1750; the first five letters on an English typewriter are QWERT. After a first reading of the complete list, the children answered twenty-five questions based directly on the information. Six trials were given, and the order of asking the questions (*not* of reading the statements) was altered each time.

(5) *Word Memory*. A list of sixteen three or four-letter words was read out to the children, and they were given two minutes to write down as many as they could remember. (In practice they were allowed as much time as they needed). This was repeated six times.

(6) *Perceptual Detail Learning*. The above tests covered the kind of material, the children learn in the verbal and number aspects of their school work. This test and the two following attempted to see how they learn using visual cues. The children were briefly shown a large card with twelve circles representing faces Each face was incomplete in some detail of eyebrows, eyes, nose, and mouth. No other features were shown, and those that were shown used only dots or straight lines; no drawing skill was involved. Having looked at the twelve faces for eight seconds, they then completed a sheet containing twelve blank circles. This test was repeated six times.

(7) *Geometrical Form Learning*. This test presented a large, rather complex, geometrical figure, built up from squares, circles, triangles, diagonals, etc., twenty distinct points in all. After looking at the figure for eight seconds, the children attempted to draw this figure. Again, six trials were given.

(8) *Symbol Maze Learning*. Some kinaesthetic learning was added to the learning of visual patterns. Each child had a printed maze, and the class was shown eighteen symbols that could *not* be crossed on the maze. After seeing the forbidden symbols for eight seconds, the maze was attempted. Six trials were involved.

(9) *Group Concept Formation Test*. Although only one test of this type was included in the battery, we feel that there is much to be learned from a battery of tests exclusively of the concept-formation type, some being individual. Bruner, Piaget, Annett (1959) and others have indicated the importance of this aspect of learning, and further studies are clearly needed. This test involved the linking of nonsense syllables with pictures. The first example showed a horse—TIX, a dog—TIX, a man—FON, and a woman—FON. The children were then shown (*a*) a baby and (*b*) a cat, and asked to write the appropriate syllables after each. There were fourteen trials in all, some being repeats, with the complexity of the concepts increasing. Concepts in the middle range included things that fly, things that make a sound, and people in uniform; and at the end of the scale—things nocturnal, physical action, and modes of transport. In the older group at least it was clear that children had benefited from experience with the early, easy, concepts when tackling the later, difficult ones.

Reference Tests' The remaining tests in the battery were as follows:

(10) A forty-item, multiple-choice, vocabulary test.
(11) A reading comprehension test, similar in form to the Watts-Vernon test.
(12) Schonell's Essential Mechanical Arithmetic.
(13) Schonell's Essential Problem Arithmetic.
(14) A twenty-item number series test.

(15) Digit memory, fourteen items ranging from three to nine digits ; this was given orally and responses were written.
(16) A spatial test involving the copying of printed figures onto a dotted space.
(17) Goodenough Draw-a-Man test.
(18) Bender-Gestalt, given as a group test with each figure shown on a large card.
(19) Overall school attainment had already been examined in 1960, and similar tests were given in the Summer of 1961 to provide a criterion of later achievement.

IV.—SCORING.

Mollenkopf and Allison seem to have used the total scores on all trials as their measure of learning. While these obviously reflect improvement, they depend very largely also on initial ability. Woodrow, on the other hand, took final-minus-initial scores, that is the absolute gain. The disadvantage here is that progress on most tasks is unlikely to be linear ; thus, there can be no guarantee that the units are equal at different parts of the scale. For example, it might be easier for those who start with a score of 10 to finish with a score of 30 than for those who start at 40 to finish with a score of 50. Stake plotted individual learning curves for each task, but eventually employed three measures for each curve, namely total success, slope or curvature (i.e., a measure of gain), and goodness of fit (i.e., regularity of progress). Since the latter measures yielded few significant correlations with any other tests, we decided to study total score, initial score in the first two trials and final score in the last two trials ; (the first and last would have yielded bigger differences, but would also be likely to be poorer in reliability). However, the amount of progress was determined by correlational analysis, as follows, not by subtracting initial from final scores.

Manning and DuBois (1958) argue that the most appropriate measure of learning is " that portion of the measure of final status which is not correlated with initial status." Thus, if we denote initial (or early) performance as A, final performance as Z, and compare these with some criterion, C, or with any other test, then the correlation of effective gain with C will be :*

$$r_{(Z-A)} = r_{ZC} - r_{AC}r_{AZ}$$

Manning and DuBois do not indicate how one should correlate gain scores on one test with those on another. But we can extend the above formula to Tests 1 and 2 as follows :

$$r_{(Z_1-A_1)(Z_2-A_2)} = r_{Z_1Z_2} + r_{A_1A_2}r_{A_1Z_1}r_{A_2Z_2} - r_{A_1Z_2}r_{A_1Z_1} - r_{A_2Z_1}r_{A_2Z_2}$$

Each of the nine learning tests yielded a total, an initial and final score in each of the two groups. With nine reference tests, 1960 attainment, 1961 estimates and 1961 attainment, this means thirty-nine variables in each group. The 1,482 product-moment correlations were kindly calculated by the London University Computer Unit,† and the gain score r's were derived from these by hand, according to the above formulae. The correlations of each measure with the two 1961 (criterion) variables were generally very similar; hence they were combined by Correlation of Sums formula. The resulting validities are shown in Table 1.

---

* Presumably this expression should be corrected, as in the ordinary formula for partial correlation, by being divided by $\sqrt{(1-r^2_{AC})(1-r^2_{AZ})}$. However, it seemed better to omit this step, since it resulted in boosting many of the Gain score correlations (because r_{AZ} was usually very large). In other words, we followed the same procedure as in calculating the residuals in a centroid factor analysis.

† These may be borrowed on request to the authors.

TABLE 1

CORRELATIONS OF LEARNING TEST GAIN AND TOTAL SCORES AND REFERENCE TESTS WITH ATTAINMENT CRITERION.

Learning Tests	Gain Scores		Total Scores		Reference Tests	8–9	10–11
	8–9	10–11	8–9	10–11			
Language Learning	·163	·201	·810	·598	Vocabulary	·567	·476
Number Substitut'n	·011	·547	·655	·787	Reading Compr. ..	·595	·666
Verbal Instructions	·256	·206	·563	·189	Mechanical Arith.	·777	·812
Verbal Information	·199	·203	·762	·688	Problem Arith. ..	·692	·657
Word Memory	·112	·276	·770	·296	Number Series ..	·639	·783
Perceptual Detail ..	·306	–·042	·517	·198	Digit Memory ..	·428	·258
Geometrical Form	·134	·361	·317	·102	Spatial Test	·599	·393
Symbol Maze	·150	·332	·465	·367	Goodenough Draw.	·382	–·117
Concept Formation	·065	·251	·535	·574	Bender-Gestalt ..	·311	·359
					1960 Attainment..	·870	·830

V.—RESULTS.

It may be seen that the validities of the Total scores are similar to those of the Reference tests, though rather variable. None is as predictive as all-round attainment at the beginning of the year. By contrast, the Gain score validities are low; yet all but one are positive. Note that there is no consistency here between the two age groups, suggesting that the tests with near-zero validities are those which were too difficult for one group to register much progress (e.g., Number Substitution for 8–9 years), or too easy for the other (Perceptual Detail for the 10–11 years). In any case the Standard Errors of correlations based on thirty-six cases are so large that inferences as to the worth of particular tests would be unsafe. For the same reason, neither multiple regression analysis nor factor analysis seemed likely to help in elucidating the results. Instead, we condensed the correlation matrices by Correlation of Sums formula to yield the correlations among groups of similar tests. These are shown in Table 2 (8–9 year above, and 10–11 year below, the diagonal).

TABLE 2

INTER-CORRELATIONS AMONG COMBINED GROUPS OF VARIABLES.

	No. of tests	Verb. Gain	Non-Verb. Gain	Verb. Total	Non-Verb. Total	Verb. Abil.	No. Abil.	Spat. Abil.	1960 Att.	Crit-erion
Verbal Gain	5		·131	·388	·264	·108	·265	·109	·289	·318
Non-Verbal Gain ..	4	·491		·219	·617	·092	·395	·260	·369	·272
Verbal Total	5	·595	·477		·590	·562	·725	·367	·796	·893
Non-V. Total	4	·411	·670	·616		·353	·599	·453	·630	·662
Verbal Ability	2	·371	·393	·539	·402		·417	·501	·670	·625
Number Ability ..	4	·622	·569	·789	·682	·608		·452	·752	·795
Spatial Ability	3	·208	·450	·219	·493	·222	·396		·526	·526
1960 Attain.	1	·442	·419	·778	·501	·700	·840	·315		·870
1961 Criterion	2	·512	·399	·702	·486	·621	·783	·285	·830	

All the gain score correlations with other variables are noticeably lower in the 8–9 year than the 10–11 year group, indicating poorer reliability and greater specificity among younger children. Even the combined Verbal Gain measures (Nos. 1-5) correlate only 0·131 with the combined Non-verbal measures (Nos. 6–9), as against 0·491 in the older group. In contrast the two sets of Total scores show very similar relations to other variables in the two groups (except to the non-verbal reference tests ; the Goodenough, Bender and Space test had very little in common among 10–11 year olds). Thus, overall performance at the new tests is quite predictive in the second year of the junior school, but children of this age do not show as consistent an ability to make steady progress at unfamiliar tasks.

Secondly, it may be noted that Gain scores tend to correlate more highly with Totals on the same tests than with any other variables. Children with good ability in any task do tend to progress more rapidly at this or similar tasks.* The same point is brought out by the *mean* correlations of the nine Gain scores with :

	8–9	10–11
Other Gain scores	·051	·141
Own Total scores	·375	·472
Other Totals	·113	·174
Reference tests and other variables	·111	·215

It would follow that rate of progress on any one task will not tell the teacher or psychologist much about performance on any other task, particularly with younger children. At the same time, Woodrow, Stake and other writers are not justified in regarding learning ability as almost wholly specific. For example, the mean inter-correlation of 0·141 among 10–11 year olds implies that our battery of Gain scores would correlate 0·60 with another similar battery ; and doubtless better tests could be devised. Moreover, a certain amount of patterning among the Gain scores may be observed in Table 2. In both groups the Non-Verbal Gains correlate more highly than do the Verbal Gains with the spatial-pictorial reference tests. However, the Totals for the learning tests show this patterning more consistently.

Prediction of Attainment. The matrices were now further condensed by combining all Gain scores, all Totals, and all Reference tests (but omitting the Perceptual Detail Learning and Goodenough in the 10–11 group). Multiple correlations with the 1961 criterion were calculated, and the beta weights were as follows :

	8–9	10–11
Gain scores combined	−·112	·093
Total scores combined	·546	−·003
Reference tests	·231	·198
1960 Attainment	·285	·613
Multiple R	·939	·845

---

* This might be expected from the derivation of the Gain scores, since the correlation of Gain with Total = $r_{ZT} - r_{AT}r_{AZ}$ and this expression can be zero or negative only if r_{AT} exceeds r_{ZT}. In the light of Fleishman's work, it seems unlikely that a person's initial performance could more closely resemble his overall success than does his final performance.

The results differ markedly for the two groups. In the 10–11's, the previous year's attainment is much the best predictor. The learning test Total adds nothing, but Reference tests cover some more ground, and Gain scores make a very small additional contibution. In the younger group the Gain scores add nothing to the other predictors ; indeed, their beta weight is negative. Previous year's attainment is somewhat less diagnostic than at 10–11, but both the Learning Test Totals and the Reference tests do contribute appreciably.

One other possibility is that the Final scores on the learning tests would not only be much more convenient than any form of Gain score but might also be the most predictive type of score. In so far as both the Total score and rate of progress must depend to some extent on the rather chancy initial reaction to an unfamiliar situation, it might be argued that the performance reached after a certain amount of practice at the new task would be a better indicator of scholastic capabilities. This is borne out among the 10–11 year pupils, the correlation of combined Final scores with the criterion being 0·705, which is higher than 0·661 for the combined Total scores. Note, however, from Table 2 that it is still below the validities of the combined Reference tests and of 1960 attainment, though it could probably be considerably improved if the more promising learning tests only were used. In the younger group the Final scores all give lower correlations than the Total scores with the criterion, though here too, some of the tests seem promising, e.g., Language Learning 0·797, Verbal Information 0·719.

VI.—Conclusions.

Our search for tests involving actual learning which might be more predictive of future scholastic achievement than 'cross-sectional' tests has proved somewhat disappointing, particularly with children in the second junior year, though more successful among older children. Apparently, such tests need to be graded carefully to the level of ability of the particular group of pupils, otherwise the amount of progress made in a limited time is highly unreliable and specific. However, it is not true, as some writers have maintained, that there is no consistency. When Residual Gain is measured, a considerable general learning factor appeared among the older children, and there was some indication of group factors for learning different types of materials—verbal, numerical, non-verbal, etc. Each Residual Gain score also correlated relatively highly with overall success at that particular task.

The predictive value of the various learning tests that were tried out varied widely, yielding the expectation that a battery of successful ones could be devised; though much more exploration with larger numbers of cases is needed. Finally, it would appear that rate of progress or gain scores possess no inherent advantage. It is the final scores achieved after instruction and practice at some unfamiliar task that yield the highest validities, and seem most likely to add to the predictive value of of conventional tests that are applied without such practice.

Useful qualitative observations can be made during the tests of pupils' methods of work, concentration, etc. However, the outlook for individual learning tests which could be applied in the educational clinic does not appear very promising, both because these would have to be very lengthy to allow adequate practice, and because a child's reaction to the learning situation in the clinic setting might differ considerably from his reaction in the classroom.

V.—References.

Allison, R. B. (1954). *Learning Measures as Predictors of Success in Torpedoman's Mates School.* Princeton, N.J. : Educational Testing Service, ONR Project 151-113.

Allison, R. B. (1956). *Learning Measures as Predictors of Success in Pipefitter and Metalsmith Schools.* Princeton, N.J. : Educational Testing Service, ONR Project 151-113.

Allison, R. B. (1960). *Learning Parameters and Human Abilities.* Princeton, N.J. : Educational Testing Service, Technical Report.

Anastasi, A. (1930). A group factor in immediate memory. *Arch. Psychol.*, No. 120.

Anastasi, A. (1932). Further studies on the memory factor. *Arch. Psychol.*, No. 142.

Anderson, J. E. (1940). The prediction of terminal intelligence from infant and pre-school tests. *Yrbk. Nat. Soc. Stud. Educ.*, 39 (I), 385-403.

Annett, M. (1959). The classification of instances of four common class concepts by children and adults. *Brit. J. Educ. Psychol.*, 29, 223-236.

Buckingham, B. R. (1921). Intelligence and its measurement. *J. Educ. Psychol.*, 12, 271-275.

Burroughs, G. E. R., and Morris, J. N. (1962). Factors involved in learning a simple musical theme. *Brit. J. Educ. Psychol.*, 32, 18-28.

Dearborn, W. F. (1921). Intelligence and its measurement. *J. Educ. Psychol.*, 12, 210-212.

Fleishman, E. A., and Hempel, W. E. (1954). Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika*, 19, 239-252.

Fleishman, E. A., and Hempel, W. E. (1955). The relation between abilities and improvement with practice in a visual discrimination reaction task. *J. Exper. Psychol.*, 49, 301-312.

Mackay, G. W. S. (1962). *A Study of the Prediction of Learning Ability in the Primary School.* M.A. Thesis, University of London Library.

Manning, W. H., and DuBois, P. H. (1958). Gains in proficiency as a criterion in test validation. *J. Appl. Psychol.*, 42, 191-194.

Melton, A. W. (1947). *Apparatus Tests.* Army Air Forces Aviat. Psychol. Prog. Res. Rep. No. 4. Washington, D.C. : U.S. Government Printing Office.

Mollenkopf, W. G. (1957). *The Gear Assembly Test : Description and Preliminary Results.* Princeton, N.J. : Educational Testing Service. ONR Project 151-113.

Ortar, G. R. (1960). Improving test validity by coaching. *Educ. Res.*, 2, 137-142.

Simrall, D. (1947). Intelligence and the ability to learn. *J. Psychol.*, 23, 27-43.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometr. Monogr.*, No. 1.

Vernon, P. E. (1961). *The Structure of Human Abilities* (2nd edit.). London : Methuen.

Woodrow, H. (1938). The relation between abilities and improvement with practice. *J. Educ. Psychol.*, 29, 215-230.

Woodrow, H. (1939). Factors in improvement with practice. *J. Psychol.*, 7, 55-70.

(*Manuscript received* 22*nd November*, 1962.)

CRITICAL COMMENTS

TEACHING EDUCATIONAL PSYCHOLOGY IN TRAINING COLLEGES*

The Report of a Joint Working Party of the British Psychological Society and the Association of Teachers in Colleges and Department of Education set up in 1959 to study the teaching of psychology to training college students and to make recommendations.

I

By T. H. B. HOLLINS

(*James Graham College, Leeds.*)

The Committee of Psychologists of the B.P.S. and the A.T.C.D.E. has produced a valuable report. It consists mainly of an analysis of the results of a questionnaire, on the teaching of psychology, sent to colleges in January, 1960, and recommendations for the future. The tone is critical without being damning and the suggestions constructive; they are not too idealistic to be practicable. The report is short and cogent; it is remarkably free from the hazy, high-sounding generalizations which afflict so much writing in Education, particularly from committees.

The picture presented here of the teaching of educational psychology in the colleges, at the end of the two-year course, is an unfavourable one. First, the proportion of unqualified lecturers teaching psychology is too big—39 per cent. (I shall speak in the present tense, as it is doubtful if the position has changed markedly since 1960): and the position is worse than this figure indicates because, of course, some colleges have a number of specialists, more than their 'share,' and too many colleges have none at all. In case it is argued that qualifications are unimportant, the report shows, from an analysis of examination papers in Education, set in 1959, of all the A.T.Os., that the teaching is unsatisfactory; some of it is out-of-date, e.g., questions on McDougall's list of instincts; most of it does not take account of recent work, e.g., there were very few questions on Piaget's contributions; and there were too many questions which could be answered merely by describing school experiences. In short, "there was little evidence . . . that students were expected to possess specific knowledge in educational psychology as they might be expected to possess it in geography or physics or English." In view of these conclusions, it is cold comfort to find that "the allocation of time to the teaching of educational psychology is reasonable"; it is about 10–15 per cent. of the whole college course, and about 40 per cent. of the Education course.

The picture would be even blacker, I think, if the analysis had gone further. A closer examination of the qualifications of specialists in educational psychology would reveal that too many of them are inadequate. Too many diploma courses allow students to qualify by theses describing children's behaviour on a basis of inadequate theory and a lack of rigorous experimental design with appropriate statistics. Wiseman† pointed out some of these weaknesses in 1959; he analysed theses and papers in educational psychology and showed how few of them were concerned with the fundamental theory of learning and child development, and how strong was the bias towards descriptive studies of behaviour. Again, I am sure the committee would have been saddened by an analysis of the *answers* to the 1959 questions, instead of the *questions*. In the first place, in my experience, the average student, answering a paper in which he has to choose four questions out of ten, avoids any question in which he may be forced to produce precise information. Secondly, he tends to give the same answer to the four he does choose, e.g., on treating maladjustment—affection, security and play; on learning—the essentials are affection, security and play; in fact, if the rubric were to forbid the mention of 'play' few would be able to answer at all, at least, amongst students training for the primary school.

---

* Copies of the Report are available from the offices of the B.P.S. at Tavistock House South, Tavistock Square, W.C.1., or the A.T.C.D.E., 151, Gower Street, W.C.1., price 5s. (post free).

† WISEMAN, S. (1959). Trends in educational psychology. *Brit. J. Educ. Psychol.*, 29, 128-135.

Some excuse can be found for the shortcomings of the teaching if one considers such factors as the small amount of time available in the two-year course, the demands made on lecturers in most colleges, which prevented adequate study, the poor intellectual calibre of some of the students (and this report remarks on our tendency to spend a disproportionate amount of our effort on the near-failures), and the gaps and weaknesses in psychological knowledge. A worthwhile psychology text is difficult (particularly to women students, whose experience of mathematics has frightened them off any information presented in tables) and if students are getting nothing from their reading (or, worse still, learning off the text by rote) we are forced to increase the practical work (in itself, an essential approach to psychology and children) to the extent of leaving out most of the theory. Brearley,* in 1957, gives an outline of a practical course for ' primary ' students and she concludes, realistically, no doubt, " Comparatively few students are ready to understand fully the tenets of the major schools of psychology, but we have greatly failed if at least some of them are not ready to study them when they leave." Collier,† in 1960, working with men students, reports " A difficulty is getting students to read effectively . . . ; it is the *quality* of reading that troubles us " and finds that theory and practice only come together in ' Study-Practice ' ; he makes a good case for a variety of methods, mainly practical, in this article, and some of his ideas are incorporated in a valuable section on methods in this report. As for psychology itself, the report rightly warns us to make " a cautious use of psychological theory " ; but this is putting it mildly. We have some certainties in the field of abilities, but in the field of learning, and personality structure and development, which are central to the teaching situation, we are among the " Two-and-Seventy jarring Sects. " In addition, nearly all the terms we wish to use in educational psychology have built-in value judgments, which lower their scientific value ; the report regrets that so few questions are set on ' motivation,' but maybe the lecturers, after reading a few pages of Peters' *Concept of Motivation*, have shown a wise discretion. Of course, the study of psychology can sometimes be as good an education as the study of literature ; both will show the complexity of life and the unlikelihood of finding easy solutions to problems ; both might produce a desire to solve the problems, nevertheless ; but few student-teachers are likely to reach this level of sophistication. Some lecturers, too, become tired of keeping up with what seem merely fashions in psychology.

However, signs of discouragement are, in fact, seldom observed. The comparative neglect of many areas of educational psychology in the training colleges, up to 1960, when the three-year course came in, was due, not to pessimism about the value of psychology, but to an optimistic concentration on an educational policy which took up such psychological contributions as reinforced the policy and neglected the rest. This policy was, and is, to make the schools child-centred. A training college educationist, especially in primary work, would see the task as persuading the 18-year old student to change from the concept of learning she would have gained in her recent education, i.e., close study of a subject, to a concept of individual children learning in a controlled but permissive environment where social and emotional learning are as important as intellectual.

In this atmosphere, gentle but insistent pressure is put on both student and child to adjust to the norm of the ' happy individual ' which seems to be the ideal of the clinical psychologist. McDougall's theories seem to give psychological respectability to this aim—all the instincts and emotions working together, in the end, under the happy jurisdiction of the master-sentiment. Freud will not do ; sex and aggression seem always about to overcome reason and civilization ; or, if repression is complete, neurosis seems inevitable ; and there is a distasteful pessimism in his dictum ' The function of education is to inhibit, suppress and forbid.' I am laughing only at the excesses of the movement and deploring only what sometimes amounts to indoctrination, because, of course, one must have admiration for a policy which has helped to produce our fine infant schools and which, even if it has not yet caught up fully with Piaget, gives our students a practical understanding of such important psychological concepts as individual differences, reading-readiness and 'learning through activity and experience'; and which produces a flow of colourful and interesting books for teaching reading and number, with efficient methods for using them.

---

* BREARLEY, M. (1957). The students' understanding of children. *Educ. for Teach.*, No. 42, 13-18.

† COLLIER, K. G. (1960). The teaching of psychology in training colleges. *Brit. J. Educ. Psychol.*, 30, 103-108.

What one wants, of course, is to retain the benefits of this crusading educational outlook and to add to it the benefits which come from the more dispassionate scientific outlook which takes a cool look at the evidence and believes in experiment. It is here that a deeper, more searching, study of psychology in the training colleges, led by a psychologist properly trained in experimental procedures, would be of value. Luckily, conditions for this to come about are more favourable than in 1960. The introduction of the three-year course will give students more time to mature and to gain a better understanding of psychology. Even more important, there is a determined effort, in the institutes and colleges, to ensure that members of staff are allowed more time for reading and research. The institutes have, for many years now, provided advanced teaching in psychology, but now many more lecturers have the time and opportunity to attend them. Then, the expansion of the colleges has meant that the large college is now the rule, rather than the exception, and this means that the appointment of a qualified psychologist in nearly every education department is a possibility. Further, as the report points out in the chapter on "The Standard of Students," the colleges "continue to attract an appreciable proportion of students who have high natural endowments" and we know that, owing to the inability of the universities to expand quickly enough, we are likely to get an increasing proportion of highly intelligent students, although the range of ability is likely to remain a wide one. It is against this background that the recommendations of the report appear particularly relevant and practicable. All of them are important, but the following seem to me the four which would contribute most to a rapid development of the sound teaching of psychology and to the growth of a scientific outlook amongst the better teachers:

(1) It is desirable that at least one member of the staff should have a degree in psychology or its equivalent in a Bachelor's, Master's or Ph.D. degree in Education, in which the main content of the course is in the field of educational psychology.

(2) . . . the teaching in educational psychology, being a co-operative task in most colleges, requires that members of staff concerned with the subject should work as a team and the fullest use should be made of the specialist knowledge of the lecturer best qualified in educational psychology.

(3) More use should be made of the facilities which exist for obtaining further qualifications through full-time and part-time study. (The facilities are summarized in Appendix 6.)

(4) We press strongly for the inclusion of more advanced optional courses in educational psychology for the abler students.

It is particularly desirable that advanced courses in psychology should be made available to the more gifted students. Although the chief reasons for this are mainly of the idealistic kind, already mentioned, it should not be overlooked that the training colleges are hoping to have their certificate recognized in some way by the universities and even to have their better students working on degree-standard courses in Education. This would link up with the growing demand to have Education, and particularly educational psychology, as a first-degree subject in universities.

Of course, no startling advances can be looked for immediately. The chief obstacles to the development of psychology as a science lie outside the ranks of psychologists and training-college lecturers. Some of these are—the indifference, or even contempt, of universities, public schools and grammar schools towards psychology, which means it will continue to be starved of its proper share of the top-level of ability, and of funds for research, for some years yet; the large classes in schools, which force the teacher to be unduly concerned with class management and keep him suspicious of a science which thrives on the variations amongst individuals ; and the unenlightened teaching of mathematics and the physical sciences in schools, particularly to girls, which results in the inability of most students to study psychology beyond the anecdotal stage. Nevertheless, this report makes useful suggestions for another step forward.

II

BY J. M. THYNE

(*Jordanhill College of Education, Glasgow.*)

The Scottish colleges of education were not involved in the ATCDE/BPS inquiry, but the Report is, of course, of professional interest to them, and to their psychology departments in particular. I comment, however, as the representative of only one, albeit the largest, of them—with a present full-time student population of over 2,500.

With the first part of the Report, most of my colleagues in the other colleges would be, I think, in general agreement. Child development (social, emotional, physical, intellectual, moral) ; the nature and conditions of learning as they relate to habits, skills, remembering, understanding, and transfer ; the psychology of the basic school subjects ; intelligence and its measurement, and applications to educational issues ; these have long been major parts of the psychology syllabus for women on the three-year course. (Our departments of psychology are concerned also with graduate students, with students holding diplomas in art, music, domestic science and other technical subjects ; but women on the three-year course form the largest group of students—over 1,600 at Jordanhill, this year—and it is to them that my comments here refer.)

What to include in the psychology course is, nevertheless, a major problem, not because it is difficult to find material relevant to the training of teachers, for it is not, but because of the difficulty of deciding what must be left out. The relevant material has to be arranged in a hierarchy of relevance, and a line has to be drawn in accord with the time available, the claims of other subjects, the abilities of the students, their own psychological maturity, and other things besides. In contrast with the Report, the line I draw excludes the training of students to apply tests, but it does include an attempt to cultivate in the student an understanding of the history, structure, and function of tests, and an ability to interpret in educational terms the results of tests which will be applied to their pupils by persons with greater expertise. This example raises the question of the criteria to be applied in selecting psychological topics, a point raised again in my concluding paragraph.

I am in sympathy with the caution expressed in the same section of the Report, though I have some reservations. As the Report says, argument by analogy from animal learning, and from highly controlled experiments on human learning, is of course, dangerous. But I, and many of my colleagues, would regard a deliberate attempt to minimise this danger as one of the fundamental tasks of the psychologist who undertakes to teach his subject to teachers. It is his special responsibility (not necessarily shared by his fellow-psychologists in universities) to ensure that what he says about learning (for example) is, in fact, valid for the kind of situation which will confront his students. And he will try to do this, not by ignoring the material in question, but by restructuring it, and testing in the school situation the results of his restructuring it. For example, analysis of human habit-forming indicates that its conditions of occurrence (if not its content) are basically the same as those made explicit in classical and instrumental conditioning. Also, it is not difficult to show that certain psycho-analytic mechanisms, such as Projection, occur within the lecture-room and the classroom. Caution is not enough ; it must lead to constructive thinking. Admittedly, the touchstone of practicality serves not only to give reality to what is taught but also to reveal the sterility of what one might otherwise have taught. It is also a humiliating revealer of one's own ignorance.

I am in complete agreement with the Report's emphasis on the need to distinguish between educational psychology and educational philosophy, but I have learned to be pessimistic about the possibility of making that distinction clear to all students. Nearly all the Scottish colleges have a department of psychology and a separate department of education (an arrangement of which I heartily approve), but this physical manifestation of segregation does not always make apparent to students what the difference in function is. To take but one rather blunt example : in many education seminars students are encouraged (very properly, I believe) to express their own opinions and develop their own judgements ; whereas in a psychology seminar they may have to learn that the opinion of a student, as contrasted with his knowledge, is of little or no value. What matters here are the facts, and the hypotheses validated by reference to them. They may have to learn also that the observation of two or three children does not provide an adequate basis for the formulation of general hypotheses about child development. (Such observation can, of course, be very useful for other purposes.)

On the central part of the Report (the actual content and forms of psychology courses) it is difficult to comment. Lists of topics, even when accompanied by tables of hours, do not answer the questions that really matter. What topics are expounded in detail and not merely mentioned in passing, or in a somewhat general way ? And just what details are emphasised ? How much of what appears on the syllabus, or is referred to in lectures, seminars, and discussions, is the student expected to know and remember ? Is the material up-to-date ? Even the analysis and classification of examination questions (which must have been a formidable task) reveals remarkably little about the above issues, though some of the general questions quoted in Appendix 4 seem to imply that some parts of the courses must themselves have contained little detail. I am tempted to suggest that more precise information might have been obtained from an analysis not only of the examination questions, but also of the students' answers. Good psychologists could scarcely object to a procedure widely used within their own discipline.

Probably the major difference between the Scottish colleges and the colleges involved in the Report relates to the qualifications of the teachers of psychology. All lecturers in psychology in the Scottish colleges of education have an honours degree of which psychology is a main subject. All are trained teachers, with experience of teaching. Some have worked in child guidance clinics. In such a context I am bound to comment that the recommendations of the Report in this connection are unduly mild, though I do appreciate many of the practical difficulties.

The problem of suitable textbooks is serious, though slowly improving. So many of the texts which are sound on psychology have little or no bearing on the work of the teacher, and so many of those which are realistic about the classroom situation contain very little psychology—as psychologists nowadays would use that word. My own experience suggests that the ' general ' textbook of educational psychology may have to dispapear from training colleges. When reasonably compact texts on specific topics (such as Memory, Learning, Child Development, Intelligence) are available, they are often much more useful than the thinly spread omnibus volume. The ' specific ' texts allow not only of relevant selection but also of specialisation.

I have not attempted to pass comments on all parts of the Report, for such an attempt would occupy more space than it would deserve. I think that the type of investigation described in the Report is valuable, and the committee responsible for it merits our thanks. Now, without in any way detracting from the value of that Report, may I suggest that a further task might profitably be undertaken, namely to discover the major problems actually faced by lecturers in psychology, and the ways in which they have been tackled ? For example, what criteria are cited in selecting psychological material for a course ? And why ? Is there an optimum order of presentation ? How do we solve the problem of breadth *versus* depth ? But it would be part of such an investigation to discover what questions should be asked.

COMMENT

ON

THE RELATION OF NEUROTICISM AND EXTRAVERSION TO INTELLIGENCE AND EDUCATIONAL ATTAINMENT

A brief comment on the controversy between J. B. Biggs and R. Lynn.

BY H. J. EYSENCK

In the controversy between Lynn and Biggs, the crucial part appears to be played by the exact status of 'anxiety.' Biggs assumes that 'anxiety' and 'neuroticism' are used interchangeably by me, but this is incorrect. Neuroticism is regarded as an inherited psycho-physical disposition, closely linked with the lability of the autonomic system, which governs a person's emotional reactivity, and may predispose him to the development of neurotic disorders under suitable circumstances. Anxiety is a conditioned fear reaction which is particularly characteristic of dysthymic neurotics, i.e., of persons who are high on the factor of neuroticism and also on the factor of introversion, which is significantly correlated with conditionability. Anxiety, therefore, is a 'mixed' concept, being related both to neuroticism and to introversion. (The position is rendered rather confused by Cattell's use of the terms 'neuroticism' and 'anxiety,' which is exactly the opposite of mine.)

My theory, which links introversion and conditionability, is not necessarily antagonistic to that of Spence, which links neuroticism, regarded as a drive, to conditionability. On the theoretical level either theory, or both, or neither, might be right. My belief that Spence's theory is inadequate is based on the experimental evidence, and may very well be mistaken; the experimentally demonstrated correlation between the Manifest Anxiety Scale and tests of introversion make all the Spence-Taylor experiments using the M.A.S. equivocal. In our own work we have never succeeded in finding correlations between neuroticism and conditionability, and neurotics as a group are more conditionable than normals *only* when they are selected in such a way that introverts are over-represented.

I agree with Biggs that a great deal more work is needed at the theoretical level, but I feel even more strongly that what is needed is a greater concentration on *experimental* studies, with particular reference to parametric investigations. The frequently contradictory results obtained by investigators testing both Spence's hypothesis and my own suggest that neither hypothesis is universally valid, but requires to be closely circumscribed with respect to its validity by detailed specification by such parameters as strength of CS and UCS, CS-UCS interval, proportion of reinforced trials and so on. Lynn's study appears to me to be a logical deduction from principles which appear to have at least some experimental support, although this support may be rather circumscribed in the manner just outlined, and although it may, of course, be possible to account for his results in other terms than those premised in his theory. This hardly detracts from the value of his important demonstration.

CRITICAL NOTICES

HUNT, J. McV. (1961). *Intelligence and Experience.* New York: Ronald Press, pp. vii+416, no price given.

BY P. E. VERNON.

This is an important book on at least two counts. First, Professor Hunt puts forward the strongest case yet made for discarding the conception of intelligence as inborn potential which is predetermined by the genes and which matures regardless of environmental conditions. Secondly, he provides the most complete exposition available in English of the whole corpus of Piaget's theories and experiments on mental development, which lead him to conclude that intelligence should be regarded as " central processes comprising strategies for processing information that develop in the course of the child's interactions with his environment."

He argues that predeterminism has been implicit in developmental psychology generally, for example in the conventional distinction between maturation and learning, and in the advice given to parents by mental hygienists not to 'push' their children's intellectual growth. So strong was its influence till recently that the common reaction to experimental demonstrations of the modifiability of measured intelligence (e.g., those of the Iowa school in the 1930's) was one of derision and emotional controversy. But with the appearance of Hebb's work on the build-up of autonomous central processes and the effects of varied environmental stimulation, Piaget's descriptions of the mechanisms of accommodation and assimilation in the progressive developments of schemata, Harlow's experiments on the acquisition of learning sets, and many other lines of evidence, the whole theoretical climate has changed. Besides describing this work in detail, the book is quite fair in recognizing contrary evidence, such as Gesell's classical co-twin control experiments on maturation, and shows how these can be reconciled with the newer theories.

Hunt does not deny that genetic factors set a limit to the complexity of what can be learned, but his interpretation of the work of modern geneticists leads him to deny that we can ever isolate and measure the genetic component either in physical or mental development. Attempts to apportion hereditary *versus* environmental variance (e.g., the often quoted 80 per cent.—20 per cent.) are misleading, since at best they only reflect the environmental pressures that have operated in the sample being studied, and tell us nothing about what would happen under different environmental conditions. Many investigations of more extreme environmental deprivations indicate far greater effects, and he goes so far as to predict that the application of a sound psychology of development could raise the average intelligence by an amount equivalent to 30 points of I.Q. In the reviewer's opinion, this forecast is unnecessarily rash; we know far more about the conditions that inhibit than those that favour the full realization of intellectual potential. And, although Piaget's work has thrown a flood of light on cognitive processes, it does little or nothing to improve our control of the even more important motivational ones.

The main key to such improvement, he claims, is the matching of environment to the current level of the child's schemata; (he admits that this principle is already realized by the good school teacher). If the child is set problems which merely exercise schemata that are already organized, he gets bored; while, if they are too far ahead of him, he is uninterested or frustrated; whereas with appropriate choice of stimuli, there should be positive pleasure in learning just as we find in the baby's acquisition of talking and walking. Circumstances which are appropriate at one stage may inhibit at another; for example, the highly protective mother may provide adequate experience in the sensorimotor stage, yet deprive her child of it in the preconceptual (1½–4 year) stage. Variation of circumstances is another important principle in the development of flexible and transferable skills.

These educational implications are among the most original parts of the book. For the most part it covers ground which is already fairly familiar to educational psychologists in Britain (though perhaps less widely known in America). As is only to be expected from a polemical treatise, the author's enthusiasm sometimes runs away with him. For example, it would have been helpful if he had given some consideration to the strong counter arguments that have been put forward by Professor Burt in a series of recent articles in the General and Statistical Journals of the B.Ps.S. Again, he probably gives too much credence to Piaget and Inhelder's claim that each stage develops out of preceding ones (e.g., that realization of constancy of quantity always precedes constancy of weight, which always precedes constancy of volume); and that a wide range of restructurings tend to occur simultaneously (e.g., that the attainment of formal thinking about scientific problems is accompanied by the revolutionary attitudes about society characteristic of adolescence). The more carefully controlled studies of Peel and Lovell, for example, indicate much more variation and inconsistency than Piaget usually admits. However, there can be no doubt that this will be in influential source book for many years, particularly since, although fairly technical an parts, it is a good deal easier to follow than Hebb's or Piaget's own writings. It is strongly recommended to advanced students and to lecturers in education and psychology.

THE APPRECIATION OF BEAUTY

VALENTINE, C. W. (1962). *The Experimental Psychology of Beauty.* London: Methuen, pp. xiv+438, 45s. net.

BY CYRIL BURT

Professor Valentine's aim has been to present a systematic survey of experiments on the appreciation of art, music, and poetry, and to determine how far the psychological differences between individuals helps to account for the wide differences in what is known as 'artistic taste.' Those who have acted as subjects in these various researches have frequently been children of school age. Hence, quite apart from the intrinsic interest of the whole problem, the observations recorded and the conclusions drawn have a direct bearing on the work of teachers.

He begins by observing that philosophers, art critics, and aestheticians have so far failed to agree on what is meant by the word 'beauty'; and he himself declines to offer any formal definition. Many, indeed, would hold with G. E. Moore that, strictly speaking, the word 'beauty,' like the word 'good,' is indefinable: the most we can do is to recognize it when we see it, and point to typical instances which may serve to indicate what precisely we have in mind. In ordinary life nearly all of us distinguish certain conscious experiences as having a common quality which we indicate by the term 'beautiful,' although in other respects they may vary widely: they may be visual, auditory, verbal, emotional, or even purely intellectual. Beauty itself seems a special form or aspect of what is technically termed 'value'—a word which is interpreted either objectively (as the older writers usually did) by saying that it denotes "something that is worthy to exist," or subjectively, by saying it is "what we prefer, like, or admire." But once again, these are merely rough, ostensive descriptions. Value itself is a 'primitive concept,' and as such cannot be defined. Yet that need not prevent scientific study. Vague as they are, the descriptions given are sufficient to show what sort of experiences we propose to study, what types of material we must therefore select, and the kind of questions we need to ask.

The characterization to which, it would seem, the experiments themselves eventually lead (so Professor Valentine tells us in his final chapter) is that the experience of beauty is distinguished by a "delighted absorption in the object itself." One thing becomes increasingly clear. "Art," as Herbert Read has emphasized, "is not necessarily beauty: that cannot be said too often or too blatantly." Nor is there any unique type of feeling that we can label *the* aesthetic emotion. The experience seems rather to be a perception or perceptual judgment of some particular quality, or combination of qualities, which is as a rule accompanied by " a harmonious blend of feelings and a certain attitude of mind." The unexpected complexity of the whole process and the astonishing number of factors that affect it are perhaps the most instructive conclusions of the entire series of investigations that Professor Valentine describes.

His book is divided into three parts—the first dealing with visual experiences, the second with auditory, and the third with verbal. In each he begins with the simplest types of stimuli, and advances progressively to the more complicated. Accordingly, after a brief introduction, his first chapter starts with experiments on simple colours, and reports the results of numerous studies carried out by himself and others on colour-preferences among infants, school children, and adults from various races and countries. So far as comparable data can be obtained, it appears that during early childhood vivid colours are everywhere strongly attractive, and that red, orange, and yellow are generally preferred to green, blue, or violet. Later, green or blue tend to predominate, and white rises high in order of preference. Here at the very outset we encounter marked individual differences, and detect the operation of a variety of factors, similar to those we meet with in experiments on more complex objects. Thus, in one set of investigations it was found that children as they grow older become more and more influenced by personal association, particularly in the case of colours that are unsaturated or 'mixed' rather than saturated and 'pure': one boy, for example, says he likes a pale faded yellow-green, "because it makes him think of leaves in Autumn"; another dislikes red, "because it reminds him of blood." These illustrate the commonest or 'associative type.' Other children refer to the 'subjective' feelings the sight of colours arouse: "this yellow dazzles the eye," "that bright red makes you feel hot all over," "the green is so restful and soothing." Somewhat rarer is the so called 'character type': one girl, for instance, speaks of scarlet as "loud and aggressive"; another of pink as "rather demure." And finally, a few among the oldest adopt an entirely 'objective' attitude, and give reasons relating solely to the colours themselves: they like an ultramarine "because it is such a pure blue," and dislike a certain green "because it has too much yellow in it to be a really good green." As Professor Valentine points out, these four types, with varying frequencies, crop up in every medium and at almost every stage. Nevertheless, so he maintains, there is sufficient uniformity to suggest, in addition to the effects of experience or temperament, the influence of an innate differential response to colours of different hues, changing spontaneously as the child matures. From single colours we turn to colours in combination; and at once we find that the mere presence of another hue can greatly modify the initial preferences. However, the appreciation of colour-harmony and the distaste for colour-clash does not appear until relatively late—six or seven years, according to most observers, though with a few individuals it emerges as early as four.

From colour we proceed to form. There is an unexpected measure of agreement over preferences for certain kinds of lines, curves, and linear shapes. Here the influence of *Einfühlung* or 'empathy' is particularly striking: the spectator, so to speak, sympathetically puts himself in the place of the lines and shapes; he feels himself moving along the graceful curve or supporting the heavy horizontal oblong. 'Subjective types' and 'character types' are thus strongly in evidence. Certain zigzags are described as 'lively' or 'playful,', others as 'harsh' or 'cruel.' Rising lines seem 'active' or 'hopeful'; falling lines 'weary' or 'sad.' And the experiments sufficiently prove, if proof were needed, that there is no single line or curve which is unanimously preferred. The recurrent search by theorists like Lomazzo, Hogarth, and Winckelmann for a so-called 'line of beauty' ended, we may remember, in the serpentine curve of a question-mark.

From the canons of Polyclitus and Vitruvius onward to the present day, numerous writers have maintained that, as in music and poetry, so in art and architecture, the secret of beauty lies in the component proportions, and that these proportions could be numerically expressed. The doctrine is most succinctly summed up in a dictum of Palladio (*I Quattro Libri dell' Architettura*, 1570): "the beauty of threshold curve, therefore, the match "diameter" was given the same intermediate value (13.5′) as for the suprathreshold curves, and "luminance" was assumed equal to the test-threshold luminance for that particular diameter. The result of this, of course, is that the left side of Equation 36 remains constant, which condition reflects the fact that the match field is at a constant "effect" level (of threshold).

As can be seen in Fig. 20, the theory predicts the data in a satisfactory manner. At the low test-threshold luminance values, enhancement takes place. This is to be expected from Equation 36, since at low B_t values the spontaneous discharge term which has B_t in its denominator (see Equations 20, 36 and 41), is relatively large; variation of $C_{t,d}$ in Equation 36, therefore, should be effective. At high B_t values, however, the spontaneous discharge term is already too small for variation over the same $C_{t,d}$ range to make much difference. Or, in speculative words, high test luminance inhibits spontaneous activity by direct and/or scattered light so that as test area increases there is little or no change in spontaneous activity. Low test luminances, however, allow spontaneous activity to be effective. At low test luminance therefore, spontaneous discharge inhibition is present and thus reduced by an expanding test area through off inhibition. Test brightness, in this way increases, or threshold luminance decreases. Such predictions are borne out in Fig. 20.

It should be noted that variation of internal test inhibition, $C_{t,t}$, is not very effective since its exponent c_3 is equal to .3010, according to Equation 39. The physiological explanation for this lack of effect is not obvious at this time. We might point, however, to the data of Boynton (1951) to support the reliability of this finding. He measured the height of the *b* wave in electroretinographic recordings as a function of the area of small test fields. He found that as area increased, the *b* wave response also increased. The interesting point, however, was that he could account for the charge in *b* wave height solely on the basis of the change in the amount of scattered light hitting the retinal surround of the test field. Presumably, changes of [illegible] of light striking within the [illegible]ers neither reduced nor increased the *b* wave response. Which is to say in our own theoretical terms, that variation of $C_{t,t}$ was not effective.

DISCUSSION AND CONCLUSIONS

Our theory set out to explain psychophysical depression and enhancement in terms of physiological inhibition occurring between on fibers (depression) and between on and off fibers (enhancement). How successful has it been?

Negative Aspects

We find that although the changes that occur (in the brightness response) in all the foregoing experiments have been generally predicted, it can be seen in Table 1 that exact prediction of the results in the experiments of Heinemann (1955) and Fry and Alpern (1953) require values of constants K_3 and K_4 quite a bit different than those for the other experiments. It is doubtful that these changes in K_3 and K_4 can be attributed to uncontrolled variables (e.g., personality) that might be operating differently in the Heinemann and Fry and Alpern experiments. It seems more likely that our theory is lacking in its development of the term S_i in Formula

37. That is, that amount by which the inducing field surrounds the test field is substantially greater in the Heinemann and Fry and Alpern experiments than in the others. Before this term is developed in detail, however, it would probably be more appropriate to first know empirically what happens to test brightness as S_i is varied.

Another consideration is that our b and c equations (37 and 38) are not rational. Nor are the O formulae (40, 41, 42) rational, even though we can point to their probable physiological bases. The improvement of this situation also awaits further experimentation and knowledge.

Positive Aspects

On the positive side of the picture, our theory has integrated in an explicit fashion all the major changes that occurred in the physiological experiments and psychophysical multiple- and single-field experiments that were considered. And except for changes in K_3 and K_4 the empirically determined constants remain relatively stable, except when, because of a particular experimental procedure or lack of control (as in Beitel, 1936) a change in constants is predicted.

Predictions

An interesting prediction is suggested by the foregoing theory. This has to do with the relationship between enhancement and spontaneous discharge and the possible dependence of spontaneous discharge upon reticular activation. According to our theory, changes in reticular activation could change the value of the term D in Equation 35 and, therefore, change the amount of enhancement present in either multiple- or single-field situations. Preliminary experiments indicate that this might be happening.

An initial experiment by Diamond is investigating the correlation between the amount of single-field enhancement (as found in Diamond's single-field area experiment above) and the amount of reticular activation, as measured by blood pressure—Lindsley (1951) suggests this as a possible measure. Our theory would predict that the higher the reticular activation as measured by blood pressure, the greater the amount of single-field enhancement. The results, taken from a sample of 30 college students, so far show a positive and significant (at the .01 level of confidence) Spearman rank order correlation between enhancement and blood pressure as predicted. The correlation is low but it still suggests support of our theory; i.e., factors other than reticular activity might be causing blood pressure differences from S to S, thus reducing the correlation. This experiment has yet to be performed for the multiple-field situation.

A second investigation by Kleman, Diamond, and Smith is studying the relationship between enhancement and the administration of a drug that might affect reticular activation. The drug that is being used is caffeine; it is a stimulant and also safe to use with college *S*s. Heinemann's (1955) multiple-field situation is employed (see Fig. 11) in which enhancement appears for the low inducing- and test-field luminances (see Fig. 12). Our initial findings, in Fig. 21, show that caffeine administration causes enhancement to disappear.

Alternative theoretical explanations seem possible at this point. A first hypothesis is that caffeine stimulates the excitatory portion of the reticular system; this stimulation is relayed to the retina, where, as a result, on fiber activity is increased more than is off fiber activity. Granit (1955, pp. 106–

107) has shown that on as well as off fiber activity increases with reticular stimulation. A relatively strong increase in on activity would weaken off activity and reduce enhancement.

A prediction accompanying this hypothesis is that caffeine administration should lower absolute threshold for a single field (since on fiber activity is supposed to increase). We find in a preliminary experiment that this does happen for some *S*s but not for others. Why these individual differences exist in this respect is now being explored. An alternative hypothesis is that caffeine acts directly on the on fibers in the retina to reduce enhancement in the above manner. A choice of the best of these two hypotheses awaits further experimentation.

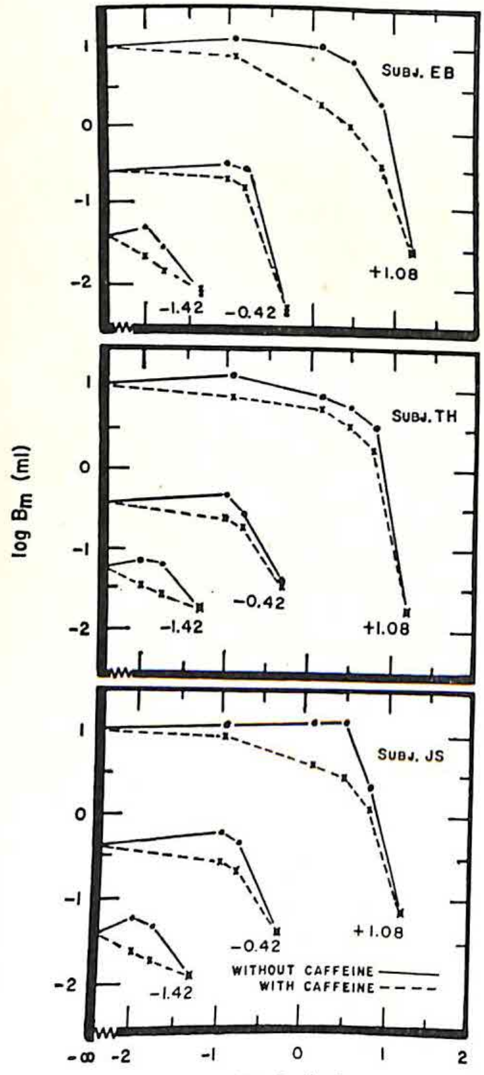


FIG. 21. The effect of caffeine on the test-brightness versus inducing-luminance function. Note that enhancement disappears with caffeine administration. (Kleman, Diamond, & Smith, unpublished.)

REFERENCES

ADRIAN, E. D. *The Basis of Sensation.* London: Christophers, 1928.

ADRIAN, E. D., & MATTHEWS, R. The action of light on the eye. Part III, The interaction of retinal neurons. *J. Physiol.*, 1928, **65**, 273–98.

ALPERN, M. Metacontrast. *J. opt. Soc. Amer.*, 1953, **43**, 648–657.

BARLOW, H. B., FITZHUGH, R., & KUFFLER, S. W. Resting discharge and dark adaptation in the cat. *J. Physiol.*, 1954, **125**, 28P.

BARTLEY, S. H. The psychophysiology of vision. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology.* New York: Wiley, 1951. Pp. 921–984.

BEITEL, R. J. Inhibition of threshold excitation in the human eye. *J. Gen. Psychol.*, 1936, **14**, 31–61.

BOYNTON, R. M., & RIGGS, L. A. The effect of stimulus area and intensity upon the human retinal response. *J. exp. Psychol.*, 1951, **42**, 217–226.

BROCA, A., & SULZER, D. La sensation lumineuse en fonction du temps. *J. physiol. path. Gen.*, 1902, **4**, 632–640.

DIAMOND, A. L. Foveal simultaneous brightness contrast as a function of inducing- and test-field luminances. *J. exp. Psychol.*, 1953, **45**, 304–314.

DIAMOND, A. L. Foveal simultaneous contrast as a function of inducing-field area. *J. exp. Psychol.*, 1955, **50**, 144–152.

FRY, G. A., & ALPERN, M. The effect of a peripheral glare source upon the apparent brightness of an object. *J. opt. Soc. Amer.*, 1953, **43**, 189–195.

GRAHAM, C. H. The relation of nerve response and retinal potential to number of sense cells illuminated in an eye lacking lateral connections. *J. cell. comp. Physiol.*, 1932, **2**, 295–310.

GRAHAM, C. H., BROWN, R. H., & MOTE, F. A. The relation of size of stimulus and intensity in the human eye: I. Intensity thresholds for white light. *J. exp. Psychol.*, 1939, **24**, 555–573.

GRANIT, R. *Receptors and Sensory Perception.* New Haven: Yale Univer. Press, 1955.

HANES, R. M. Supra-threshold area brightness relationships. *J. opt. Soc. Amer.*, 1951, **41**, 28–31.

HARTLINE, H. K. The response of single optic nerve fibers of the vertebrate retina. *Amer. J. Physiol.*, 1938, **121**, 400–415.

HARTLINE, H. K. The effects of spatial summation in the retina on the excitation of the fibers of the optic nerve. *Amer. J. Physiol.*, 1940, **130**, 700–711.

HARTLINE, H. K., & GRAHAM, C. H. Nerve impulses from single receptors in the eye. *J. cell. comp. Physiol.*, 1932, **1**, 277–295.

HARTLINE, H. K., & RATLIFF, F. Inhibitory interaction of receptor units in the eye of *Limulus. J. gen. Physiol.*, 1957, **40**, 357–376.

HARTLINE, H. K., WAGNER, H. G., & RATLIFF, F. Inhibition in the eye of *Limulus. J. gen. Physiol.*, 1956, **39**, 651.

HEINEMANN. E. G. Simultaneous brightness induction as a function of inducing- and test-field luminances. *J. exp. Psychol.*, 1955, **50**, 89–96.

LINDSLEY, D. B. Emotion. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology.* New York: Wiley, 1951. Pp. 473–516.

LEIBOWITZ, H., MOTE, F. A., & THURLOW, W. R. Simultaneous contrast as a function of separation between test and inducing fields. *J. exp. Psychol.*, 1953, **46**, 453–456.

POLYAK, S. L. *The Retina.* Chicago: Univer. Chicago Press, 1941.

(Received July 10, 1959)

Psychological Review
Vol. 67, No. 3, 1960

A TWO-FACTOR THEORY OF INHIBITION[1]

MERRELL E. THOMPSON
New Mexico State University

Perhaps no other aspect of Hull's (1943) learning theory has been so thoroughly explored, both empirically and theoretically, during recent years as his theoretical treatment of extinction and related phenomena. No fewer than 12 theoretical papers dealing with this problem have appeared in the *Psychological Review* during the past seven years (Cotton, 1955; Ellis, 1953; Estes, 1955; Glanzer, 1953a; Gleitman, Nachmias, & Neisser, 1954; Johdai, 1956; Maatsch, 1954; Razran, 1956; Saltz, 1953; Walker, 1958; Wasserman, 1952; Wolpe, 1952).

Gleitman, et al. (1954) presented a critical and important evaluation of Hull's formulation in which they pointed out several paradoxical consequences which can be derived from Hull's postulates, one being that the performance curve should rise to a maximum and with continued reinforced trials eventually decline to its starting point; another being that reconditioning to the original level would be impossible. Most of the criticisms of these authors were concerned with *conditioned inhibition,* a concept which many psychologists have held in low esteem for years.

The most serious attempt at formulation of an alternative theory has been made by Glanzer (1953a), who assumed that with continued exposure to the same stimulus the organism becomes less active, i.e., stimulus satiated. Glanzer makes the additional assumption that this "boredom-like" effect dissipates with time. From this theory are deduced spontaneous alternation and related phenomena.

A critical evaluation of Glanzer's formulation, together with numerous empirical studies, illustrates the limitation of the stimulus satiation concept. Glanzer, like Hull, can explain certain empirically known facts, but finds it difficult to handle others. For instance, it is difficult to see how Glanzer's formulation can explain such a phenomenon as an increased proportion of left turns following a forced right turn and an increase in this effect following two forced right turns (Grosslight & Ticknor, 1953; Thompson, 1952). Glanzer (1953a) also failed to verify Deduction 5 which appears to be basic to his theory. Deduction 5 states: "In the simple two-alternative situation, as stimuli differentiating the alternatives are eliminated, spontaneous alternation will decrease" (p. 262).

In their treatment of inhibitory potential both Hull and Glanzer neglect several variables which may have considerable influence upon the development and dissipation of inhibition. The diverse results of the reported studies appear to be complicated by differences in such variables as the nature of the first trial in a block of two consecutive trials (free or forced) and reward for both alternatives or no reward in the maze. Both of these authors failed to differentiate between reinforced and nonreinforced responses on inhibitory potential. The majority of studies cited in support of the stimulus satiation theory appear to employ nonreinforced responses. If this variable should lead to results which differ from studies using reinforced responses, we might find that both Hull's reactive inhibition postulate and Glanzer's stimulus satiation postulate have merit.

---

[1] This work has been supported in part by Grant G-5817 from the National Science Foundation. An abbreviated version of the present paper was given at the 1958 meeting of the Rocky Mountain Psychological Association. Appreciation is expressed to Jean P. Thompson for valuable assistance in preparing this paper. Thanks are also due Joseph H. Forsyth for critically reading the manuscript.

In view of the conflict between empirical data and the deductions from existing theories of inhibitory potential, a need is indicated for a more comprehensive theory than the one presented by Glanzer, one which also avoids the paradoxes derived from Hull's formulation.

The two-factor theory of inhibitory potential presented in this paper was first designed in an attempt to explain some of the apparently conflicting results of studies on spontaneous alternation. The theory is a modification of Hull's reactive inhibition postulate and Glanzer's stimulus satiation postulate. This revised formulation should not only enable us to take care of some of the contradictions in the theories presented by Hull and Glanzer, but yield novel predictions, and help to explain empirical results from a number of apparently unrelated areas.

Following the procedure of Hull and Glanzer, the theory is presented in the form of three postulates, as follows:

Postulate 1. Whenever an organism makes a response,[2] there is developed an increment of inhibition which decreases the probability of recurrence of that same response. This condition is called response inhibition (I_R). The amount of I_R developed by a series of response evocations is a negative growth function of the number of evocations, and dissipates as a simple negative growth function of time.

Postulate 2. Whenever an organism makes a response, or is exposed to a stimulus in the absence of primary reinforcement, there is developed an increment of inhibition which decreases the probability that the organism will return to or respond to the same stimulus, providing the organism has previously been rewarded in the presence of similar stimuli (e.g., in previous maze training or during preliminary training). This condition is called stimulus inhibition (I_S). The amount of I_S developed is a negative growth function of number of stimulus exposures. Stimulus inhibition dissipates very slowly with the passage of time and will generalize from one stimulus to other similar stimuli.

Postulate 3. Response inhibition and stimulus inhibition summate functionally to produce total inhibition (I_T).[3] Implicit in the postulates are the following statements:

1. Whenever a response is made and followed by reinforcement, only I_R will develop.

---

[2] The term response as used here refers to acts producing changes in the immediate environment or in the relation of the organism to the immediate environment. Examples of response as defined above would be turning right in a maze, approaching a food cup, pressing a bar, running down a maze alley, etc.

[3] Taking the modification of Hull's theory (as presented by Spence, Goodrich, & Ross, 1959) as a point of departure the above three inhibitory constructs could be incorporated as follows:

In a simple two-alternative situation when one alternative is rewarded, the excitatory potentials of the responses to the two discriminanda (S_+S_-) are given by the following equations:

$$E_+ = H_+ \times (D + K) - I_R$$
$$E_- = (H_- \times D) - I_T$$

where $H = fN_R$ (number of trials run), $D = fTd$ (time of food deprivation), $K = fWg$ (magnitude of reward), $I_R = fN_R$, $I_S = fn$ (number of nonreinforced trial runs), and $I_T = I_R + I_S$.

2. Whenever an organism is exposed to a stimulus in the absence of reward, only I_S will develop. This is based on the assumption that the organism is not required to make a specific response in arriving at the stimulus.

3. Whenever a response is made and not followed by reinforcement, both I_R and I_S will develop.

4. When responses are learned under conditions of partial reinforcement, both types of inhibition will develop, I_R on every trial and I_S on the nonreinforced trials.

The postulates yield a large number of deductions which are relevant to several experimental situations. Some of the more basic of these are formulated and discussed below. They are divided for convenience into two main categories: deductions relevant to spontaneous alternation, and deductions relevant to extinction and related phenomena. The former are further divided into situations of constant stimuli and situations of reversed stimuli.

DEDUCTIONS RELEVANT TO SPONTANEOUS ALTERNATION

In a simple two-alternative situation, such as a T or Y maze, where the subject (*S*) is started from the same starting box on two consecutive trials, thus maintaining constant stimuli, the following deductions apply:

Deduction 1: Nonrewarded *S*s will alternate more than rewarded *S*s.

Since only I_R is developed when a response is followed by reward, and both I_R and I_S develop when a response is not rewarded, it would follow that the total amount of inhibition against repeating the turn and approaching the stimulus would be greater in the latter case. A verification of this deduction was made by Thompson[4] who found that nonrewarded *S*s showed significantly more alternation (73%) than rewarded *S*s (41%).

Deduction 2: There will be a decrease in alternation with an increase in intertrial interval. This decrease will be greater for rewarded *S*s than for nonrewarded *S*s.

The two-factor theory assumes that I_R dissipates much more rapidly than I_S. Therefore rewarded *S*s will show greater decrease in alternation after a passage of time since all of the inhibition will dissipate, whereas the nonrewarded *S*s will lose their I_R but less of the I_S, thus showing less decrease in alternation after a passage of time.

Solomon (1948) and Dember & Fowler (1958) cite many studies verifying the first part of this deduction. The second part of the deduction was verified by Thompson (see Footnote 4) who found that nonrewarded *S*s showed a 16% decrease in alternation with 30-min. intertrial interval while rewarded

---

[4] Eighty rats were run on a cross-shaped rotating maze with two starting boxes in an attempt to study alternation as a function of three variables: (*a*) similar (black-black) vs. dissimilar (black-white) alleys and end boxes, (*b*) reinforcement in both end boxes vs. no reward in either end box, and (*c*) the first trial in each block of two being forced or free. By using a 2 × 2 × 2 factorial design it was possible to determine the relative importance of each of the three main variables as well as possible interactions.

Each *S* was given four trials, in blocks of two, each day for 12 days. The two trials within each block were given with a 30-sec. intertrial interval and the interblock interval was 30 min. On half of the 24 blocks the *S*s were given their two trials from the same starting box, either S_1 or S_2, thus maintaining a constant position of maze arms. On the other 12 blocks, the *S*s were started in the S_1 (or S_2) box for trial one and then placed in the other box for trial two, thus reversing the relative position of the maze arms.

*S*s showed a 61% decrease with the same intertrial interval.

In a simple two-alternative situation, such as a cross-shaped maze, when the *S* is started from opposite starting boxes on two consecutive trials, thus reversing the stimuli, the following deductions apply:

Deduction 3: Rewarded *S*s will alternate turns and repeat stimuli.

This deduction follows from the assumption that only I_R develops when responses are rewarded. Since no I_S is developed, there will be no tendency for the *S*s to alternate stimuli. This implies that Hull's reactive inhibition postulate is essentially correct if applied only to situations in which both alternatives are rewarded. This deduction has not been verified.

Deduction 4: Nonrewarded *S*s will alternate stimuli and repeat turns when the maze arms are dissimilar.

The development of I_R would lead to the prediction that the *S*s would alternate turns. However, I_S will develop and accumulate and eventually overcome the effect of I_R; i.e., the total amount of I_S after a few trials will be greater than the amount of I_R, which will dissipate between blocks of trials. Glanzer (1953b), using a black-white maze found 81% alternation of maze arms (stimuli) when his *S*s were started from opposite starting boxes on two consecutive trials. Thompson (see Footnote 4) found nonrewarded *S*s run on the black-white maze showed 60% alternation of stimuli.

Deduction 5: Nonrewarded *S*s will alternate turns and repeat stimuli when the arms are similar.

Since I_S is assumed to generalize to similar stimuli, the I_S developed to Arm A will generalize to Arm B and since the amount of I_S is approximately equal for both of the alternatives, the *S*s response will be determined largely by I_R. Thompson (see Footnote 4) found verification of this deduction in the fact that nonrewarded *S*s run on the black-black maze alternated turns 65% of the time as compared to only 40% turn alternation on the black-white maze.

Deduction 6: Rewarded *S*s will show a decrease in turn alternation and an increase in alternation of stimuli with an increase in intertrial interval.

The stimulus satiation theory makes the opposite prediction, i.e., a decrease in alternation of stimuli with an increase in intertrial interval since Glanzer assumes that stimulus satiation effects dissipate with time. The two-factor theory, assuming that only I_R develops when the responses are reinforced, would predict a decrease in turn alternation as the I_R dissipates. Since the *S* can alternate turns *or* stimuli but not both, any decrease in turn alternation must lead to an increase in the alternation of stimuli. Verification of this deduction was found in Thompson's (see Footnote 4) study; the percentage of turn alternation for 30-sec., 30-min., and 24-hr. distribution was 40, 29, and 28 respectively.

DEDUCTIONS RELEVANT TO EXTINCTION, REMINISCENCE, AND SPONTANEOUS RECOVERY

Deduction 7: Extinction, as measured by a decrease in spontaneous recovery, can be strengthened by continuing nonreinforced presentation of the CS after the CR has been extinguished to a point of nonelicitation.

This is sometimes referred to as *below zero extinction* and has been verified by Pavlov (1927) and Brogden, Lipman, & Culler (1938). According to the two-factor theory, both I_R and I_S develop during the extinction trials up to the point of nonelicitation of the CR with I_R dissipating after this point has been reached. Continued exposure of the *S* to the stimulus (CS) in the absence of reward would produce still more I_S which leads to a predicted decrease in spontaneous recovery.

Deduction 8: After rats have been trained to go to one side of a T maze, the number of trials to extinguish the response may be reduced by placing the *S* in the previously rewarded goal box without reward, prior to extinction proper.

According to the two-factor theory, I_S develops when the *S* is exposed to the stimulus in the absence of reward (as during latent extinction). Thus the number of extinction trials required are reduced. This deduction agrees with the experimental findings of Coate (1956), Deese (1951), Denny & Ratner (1959), Hurwitz (1955), Moltz (1955), Moltz & Maddi (1956), Seward & Levy (1949), and Thomas (1958).

Deduction 9: Spontaneous recovery will be reduced when latent extinction precedes experimental extinction, provided control and experimental *S*s are given the same number of extinction trials.

The amount of I_R developed during experimental extinction is approximately the same for the two groups. However, the experimental group has more I_S, since some I_S was developed during the latent extinction period. Since I_S is assumed to dissipate at a very slow rate, the amount of spontaneous recovery for the experimental group should be reduced. Thomas (1958) found no significant difference between groups in amount of recovery, but his experimental design was such as to offer no direct test of the above deduction.

Deduction 10: When secondary reinforcement is minimized, continuously reinforced *Ss* will show superior performance to partially reinforced *S*s at all stages of acquisition when the two groups are compared in terms of number of reinforced trials. The superiority of the 100% group will decrease with increase in intertrial interval.

The habit strength of both groups is equal since they have had the same number of reinforced trials. However, the amount of total inhibition is greater for the partially reinforced group since the total number of trials exceeds that of the 100% group. With an increase in intertrial interval the I_R dissipates for both groups, leaving the I_S developed on the nonreinforced trials for the partially reinforced group and thus decreasing the difference in performance. Verification of this deduction is seen in a recent study by Reynolds (1958) who found that when performance was compared in terms of number of reinforced trials, the curve for the 100% reinforced group remained significantly above that for the 60% reinforced group.

Deduction 11: The extinction curve of a continuously reinforced group will have a steeper initial slope than that of a partially reinforced group.

This deduction is based on the assumption that I_S develops according to a negative growth function of number of nonreinforced trials and dissipates very slowly. Since the partially reinforced group will have some I_S developed during the nonreinforced acquisition trials, the increment of I_S per trial will be smaller during the early extinction trials than for a continuously reinforced group. Verification of Deduction 11 is found in many studies, among them James & Rotter (1958), Reynolds (1958), Sheffield (1949), Weinstock (1954, 1958), and Wilson, Weiss, & Amsel (1955).

Deduction 12: Reminiscence will be complete for *S*s trained under continuous reinforcements; i.e., following rest, the performance curve for a massed group will be the same as the curve for a distributed group.

Since only I_R is assumed to develop on reinforced trials, reminiscence should be complete as I_R dissipates rapidly with the passage of time. Several studies on motor learning (Adams & Reynolds, 1954; Archer, 1954; Bilodeau, 1952; Schucker, Stevens, & Ellis, 1953; Starkweather & Duncan, 1954) have failed to verify Kimble's (1949) reported evidence for permanent work decrement on postrest performance following massed acquisition trials.

Deduction 13: Reminiscence will be incomplete for a long period of time following acquisition under partial reinforcement conditions; i.e., following rest, the performance curve for a partially reinforced group will be below the curve for a continuously reinforced group, provided the number of reinforced trials is held constant and secondary reinforcement is minimized. The amount of this "permanent" work decrement will be proportional to the number of nonreinforced trials.

This deduction is based on the assumption that I_S developed on the nonreinforced trials will dissipate very slowly, thus prolonging complete recovery of the performance curve.

Deduction 14: Spontaneous recovery will be less for a group which learned the response under partial reinforcement conditions than for a group which learned under continuous reinforcement, provided both groups are given the same number of extinction trials.

Total inhibition developed during extinction is the same for both groups, but the I_S resulting from the nonreinforced acquisition trials of the partially reinforced group yields a total amount of I_S exceeding that of the continuously reinforced group and results in less spontaneous recovery.

The author knows of no studies directly relevant to Deductions 13 and 14. However, several studies testing these and other deductions relevant to reminiscence and spontaneous recovery are underway in our laboratory.

SUMMARY

Because both the reactive inhibition and stimulus satiation explanations of spontaneous alternation seem inadequate, a two-factor theory of inhibition, in the form of three postulates, is presented. Deductions for the simple two-alternative situation, as well as extinction, reminiscence, and spontaneous recovery phenomena are listed, together with available supporting evidence.

REFERENCES

ADAMS, J. A., & REYNOLDS, B. Effect of shift in distribution of practice conditions following interpolated rest. *J. exp. Psychol.*, 1954, **47**, 32–36.

ARCHER, E. J. Postrest performance in motor learning as a function of prerest degree of distribution of practice. *J. exp. Psychol.*, 1954, **47**, 47–51.

BILODEAU, E. A. Performance decrement in a simple motor task before and after a single rest. *J. exp. Psychol.*, 1952, **43**, 381–390.

BROGDEN, W. J., LIPMAN, E. A., & CULLER, E. The role of incentive in conditioning and extinction. *Amer. J. Psychol.*, 1938, **51**, 109–117.

COATE, W. B. Weakening of conditioned bar-pressing by prior extinction of its subsequent discriminated operant. *J. comp. physiol. Psychol.*, 1956, **49**, 135–138.

COTTON, J. W. On making predictions from Hull's theory. *Psychol. Rev.*, 1955, **62**, 303–314.

DEESE, J. The extinction of a discrimination without performance of the choice response. *J. comp. physiol. Psychol.*, 1951, **44**, 362–366.

DEMBER, W. N., & FOWLER, H. Spontaneous alternation behavior. *Psychol. Bull.*, 1958, **55**, 412–428.

DENNY, M. R., & RATNER, S. C. Distal cues and latent extinction. *Psychol. Rev.*, 1959, **66**, 33–35.

ELLIS, D. S. Inhibition theory and the effort variable. *Psychol. Rev.,* 1953, **60,** 383–392.

ESTES, W. K. Statistical theory of distributional phenomena in learning. *Psychol. Rev.,* 1955, **62,** 367–377.

GLANZER, M. Stimulus satiation: An explanation of spontaneous alternation and related phenomena. *Psychol. Rev.,* 1953, **60,** 257–268. (a)

GLANZER, M. The role of stimulus satiation in spontaneous alternation. *J. exp. Psychol.,* 1953, **45,** 387–393. (b)

GLEITMAN, H., NACHMIAS, J., & NEISSER, U. The S-R reinforcement theory of extinction. *Psychol. Rev.,* 1954, **61,** 23–33.

GROSSLIGHT, J. H., & TICKNOR, W. Variability and reactive inhibition in the meal worm as a function of determined turning sequences. *J. comp. physiol. Psychol.,* 1953, **46,** 35–38.

HULL, C. L. *Principles of behavior.* New York: Appleton-Century, 1943.

HURWITZ, J. M. B. Response elimination without performance. *Quart. J. exp. Psychol.,* 1955, **7,** 1–7.

JAMES, W. H., & ROTTER, J. B. Partial and 100% reinforcement under chance and skill conditions. *J. exp. Psychol.,* 1958, **55,** 397–403.

JOHDAI, K. A field theory of extinction and spontaneous recovery. *Psychol. Rev.,* 1956, **63,** 243–248.

KIMBLE, G. A. Performance and reminiscence in motor learning as a function of the degree of distribution of practice. *J. exp. Psychol.,* 1949, **39,** 500–510.

MAATSCH, J. L. Reinforcement and extinction phenomena. *Psychol. Rev.,* 1954, **61,** 111–118.

MOLTZ, H. Latent extinction and the reduction of secondary reward value. *J. exp. Psychol.,* 1955, **49,** 395–400.

MOLTZ, H., & MADDI, S. R. Reduction of secondary reward value as a function of drive strength during latent extinction. *J. exp. Psychol.,* 1956, **52,** 71–76.

PAVLOV, I. P. *Conditioned reflexes.* London: Oxford Univer. Press, 1927.

RAZRAN, G. Extinction re-examined and re-analyzed: a new theory. *Psychol. Rev.,* 1956, **63,** 39–52.

REYNOLDS, W. F. Acquisition and extinction of the conditioned eyelid response following partial and continuous reinforcement. *J. exp. Psychol.,* 1958, **55,** 335–341.

SALTZ, E. A single theory for reminiscence, act regression, and other phenomena. *Psychol. Rev.,* 1953, **60,** 159–171.

SCHUCKER, R. E., STEVENS, L. B., & ELLIS, D. S. A retest for conditioned inhibition in the alphabet-printing task. *J. exp. Psychol.,* 1953, **46,** 97–102.

SEWARD, J. P., & LEVY, N. Sign learning as a factor in extinction. *J. exp. Psychol.,* 1949, **39,** 660–668.

SHEFFIELD, VIRGINIA F. Extinction as a function of partial reinforcement and distribution of practice. *J. exp. Psychol.,* 1949, **39,** 511–526.

SOLOMON, R. L. The influence of work on behavior. *Psychol. Bull.,* 1948, **45,** 1–40.

SPENCE, K. W., GOODRICH, K. P., & ROSS, L. E. Performance in differential conditioning and discrimination learning as a function of hunger and relative response frequency. *J. exp. Psychol.,* 1959, **58,** 8–16.

STARKWEATHER, J. A., & DUNCAN, C. P. A test for conditioned inhibition in motor learning. *J. exp. Psychol.,* 1954, **47,** 351–356.

THOMAS, A. R. Some variables affecting latent extinction. *J. exp. Psychol.,* 1958, **56,** 203–212.

THOMPSON, M. E. Reactive inhibition as a factor in maze learning: III. Effects in the human stylus maze. *J. exp. Psychol.,* 1952, **43,** 130–133.

WALKER, E. L. Action decrement and its relation to learning. *Psychol. Rev.,* 1958, **65,** 129–142.

WASSERMAN, H. N. A unifying theoretical approach to motor learning. *Psychol. Rev.,* 1952, **59,** 278–284.

WEINSTOCK, S. Resistance to extinction of a running response following partial reinforcement under widely spaced trials. *J. comp. physiol. Psychol.,* 1954, **47,** 318–322.

WEINSTOCK, S. Acquisition and extinction of a partially reinforced running response at a 24-hour intertrial interval. *J. exp. Psychol.,* 1958, **56,** 151–158.

WILSON, WILMA, WEISS, ELIZABETH J., & AMSEL, A. Two tests of the Sheffield hypothesis concerning resistance to extinction, partial reinforcement, and distribution of practice. *J. exp. Psychol.,* 1955, **50,** 51–60.

WOLPE, J. The formation of negative habits: A neurophysiological view. *Psychol. Rev.,* 1952, **59,** 290–299.

(Received August 31, 1959)

VOL. 67, NO. 4 JULY 1960

THE PSYCHOLOGICAL REVIEW

LEARNING THEORY AND THE NEW "MENTAL CHEMISTRY"[1]

W. K. ESTES

Indiana University

Hovering in the background of our scientific enterprises is a question which we, as investigators, can afford neither to raise very often nor to overlook entirely. Namely, do we have good reason to believe that the general methods and working assumptions underlying our research can be counted on to lead in the long run to satisfactory interpretations of our phenomena? In the psychology of learning, there has been a division of labor, with the experimentalists doing the overlooking, and each theorist having his turn at bringing up this question along with his answer. The answer is, of course, in each instance that the particular theorist's tactics point down the high road while those of the opposition lead into dark and forbidding *culs-de-sac.*

For my own part, although my hands are not entirely clean in the matter of theorizing, I had been content until very recently to go along with the single-minded "learning experimentalist," assuming that the Lord will look after those who remember their control groups and mind their Ps and Fs. The incident which jarred me out of this comfortable way of life began as a simple (in fact, as will be seen, unusually simple) bit of experimentation with no philosophical overtones. The original purpose of the experimentation was to provide more cogent empirical support than had hitherto been available for one of the central concepts of learning theory.

Several recent reviewers (e.g., Deese, 1958; Estes, 1956; Kendler, 1959; Restle, 1959) have noted that despite the popular stereotype of "learning theory" as virtually synonymous with "controversy," there has steadily, although unobtrusively, accumulated a body of concepts and assumptions which command relatively wide agreement and which contribute motivation and direction to a great part of the research being done in the field of learning. At the center of this core of communality one finds a concept which represents the distillation of centuries of theorizing about learning, not to speak of 70-odd years of experimentation in the tradition of functionalism and, later, behaviorism: the concept of associative strength. In the "mental mechanics" and "mental chemistry" of pre-experimental associationism, this concept was verbalized in terms of the strength of associations between ideas; with Thorndike, it became strength of stimulus-response bonds; with Hull, the basic quantitative constructs of habit strength and excitatory potential. And it is in terms of this concept that contemporary learning theorists express the basic distinction between learning and performance, as well as their fundamental postulates relating the growth of habits to the number of training trials ("reinforcements").

---

[1] This paper comprises, in substance, the writer's Presidential Address to the Division of Experimental Psychology, American Psychological Assocation, 1959. Several indebtednesses deserve acknowledgment: an unrestricted grant from the Ford Foundation supported portions of the experimental work; periods of relative freedom from academic routine in stimulating surroundings were made possible by the Department of Psychology, Northwestern University and the Institute for Mathematical Studies in Social Sciences, Stanford University during the spring and summer, respectively, of 1959 while the paper was in preparation.

There is little disagreement even in the precise quantitative form of these postulates, perhaps because our standards of acceptability come directly from the observed forms of empirical curves relating probabilities of learned responses, as well as measures of resistance to extinction or forgetting, to the number of training trials. Thus in a whole array of contemporary theories we find exactly the same equations expressing the effect of a reinforced training trial upon associative strength (cf. Restle, 1959). In Hull's system, this equation is

$$\Delta H = k(M - H)$$

where ΔH represents the change in habit strength (H) on any reinforced trial, M is the maximum value of habit strength, and k is a constant. Spence's (1955) revision of Hull's theory is based on the same postulate, although there is some question as to whether the H in the equation should be replaced by E, representing excitatory potential (Spence, 1955, 1958). Similarly, in statistical learning theory (Estes, 1959) the corresponding assumption has been expressed in the form

$$\Delta p = \theta(1 - p)$$

where Δp represents the change in the proportion of stimulus elements connected to a given response, and θ is a constant (the proportion of stimulus elements sampled on the trial). In the stochastic model of Bush and Mosteller (1955), the same linear function is assumed, "strength" being simply identified with response probability. For application to learning theorists, a well-known quotation may aptly be reversed to read, "In strength there is union." And with respect to the experimental literature, it seems fair to say that the assumptions that associative strength increases with reinforcement, decreases during retention intervals, and generalizes to new stimuli are the cornerstones of most contemporary treatments of conditioning and elementary verbal learning.

Areas of substantial agreement concerning either facts or interpretations, let alone both, are rare enough in the field of learning so that one might think we would do well to treasure the one we have located and carefully avoid doing anything to disturb it. This I (along with most of my fellow theorists and experimentalists) was happy to do, until my hand was forced by some purely experimental developments. The first of these was the, by now well-known, work of Kimble and his associates (1955, 1956; Dufort & Kimble, 1958) who, working with an eyelid conditioning situation, tried the novel procedure of omitting the CS on a substantial block of trials during an acquisition series. The rather surprising result of this variation was that the course of acquisition was virtually unaffected. Their interpretation was that only the first few reinforcements are actually effective in modifying the strength of associations between CS and CR, the rest of the "conditioning curve" actually reflecting some nonassociative process. The second development was contributed by Rock (1957), who found no retardation in the speed of paired-associate learning when he introduced the device of replacing missed items with new ones at the end of every trial. Rock interpreted his findings as indicating that associations form on an all-or-none basis and that it is only associative strength in the sense of resistance to forgetting that grows as a function of number of reinforcements.

It seemed curious that both of these deviations from the usual experimental paradigms had yielded results which are in some respects sharply at variance with the generally accepted conception of the acquisition process. Of course these findings might turn out to have special explanations that would leave the established conception undisturbed. But then again, they might not. In the case of the Kimble, Mann, and Dufort (1955) study, counterexplanations and counterexperiments have already begun to pour into the literature (see, e.g., Goodrich, Ross, & Wagner, 1957) with the result so far of complicating matters still further rather than of clearing anything up. One might argue that concepts and assumptions which have been supported by a large accumulation of experimental findings cannot be seriously threatened by one or two apparently negative results. But this argument is weakened by the observation that all of the empirical support for the conceptualization of learning in terms of associative strength is quite indirect in character.

The basic concepts and assumptions of learning theory are universally supposed to refer to states and processes of the individual organism. Yet the existing evidence for the assumption that associative strength is an increasing function of number of reinforcements comes from performance curves representing average response measures over groups of learners, or from measures of resistance to extinction or retention scores averaged over groups of learners having different values of the independent variable. Even the few bits of negative evidence are indirect, depending on performance curves obtained under deviations from the usual experimental paradigms but still representing changes in average scores over series of trials for groups of subjects (*S*s). It would seem that if our basic conceptions are sound, it should be possible to cut through the web of group performance curves and obtain more direct and compelling evidence for the existence of the assumed states and processes in individual organisms. This, in any event, is what I set out to accomplish for the concept of associative strength in a series of experiments now to be reported.

ON THE DEFINITIONS OF "REINFORCEMENT," "TEST TRIAL," AND "LEARNING"

In standard human learning situations, "learning" is almost universally defined and measured in terms of a change in the probability, or frequency, with which a given stimulating situation evokes a response (or instances of a response class) that has been designated as "correct" by the experimenter. With one reservation, to be noted below, I shall follow this usage. But the situation is quite different with "reinforcement," the same term being used in at least two quite different senses by different investigators and thus promoting no end of confusion. My own habitual usage is the "neutral definition" (Hilgard, 1956, p. 409) which identifies reinforcement empirically with the operation that is supplied by the experimenter in order to produce learning, as defined above, in any given situation. In a paired-associate situation, the reinforcing operation is the paired presentation of the stimulus and response members of an item; in classical conditioning it is the paired presentation of CS and US; in verbal conditioning, the reinforcing operation for a given predictive response (e.g., predicting that the left light will appear) is the occurrence of the corresponding event (appearance of the left light)—in each case without regard to whether the *S* correctly anticipated the response member of the paired-associate item, gave a CR prior to occurrence of the US, or correctly predicted the event on the trial in question. The only property that different types of reinforcing operations are assumed to share is their common quantitative effect on the conditional probabilities of the possible alternative responses to the stimulating situation in which reinforcement occurs.

A narrower definition, favored especially by writers associated with a drive-reduction interpretation of reinforcement, would limit the term reinforcement to an operation that follows and is contingent upon the occurrence of the reinforced response on any trial. In this usage, reinforcement in paired-associate learning occurs only when the *S* has made a correct response in anticipation of the paired stimulus-response presentation, and reinforcement in verbal conditioning occurs only on trials when the *S* correctly predicts the trial outcome. Whether, according to this view, reinforcement occurs on only those trials of a classical conditioning experiment on which a CR occurs prior to the US depends upon theoretical decisions as to whether the CR and UR are "the same response" and whether reinforcement occurs at the onset or the termination of the US.

The primary advantage I see in the "neutral definition" is that it can be applied in an objective and consistent manner independently of one's position on systematic or theoretical issues. Learning certainly may occur prior to the first correct anticipation in a paired-associate experiment, prior to the first correct prediction in verbal conditioning, prior to the first CR in classical conditioning. The present usage permits us to speak, for example, about changes in probability of a response as a function of reinforcements on trials preceding its first occurrence, on the one hand, and changes as a function of reinforcements on trials including and following its first occurrence, on the other, without changing our definition of reinforcement.

It should be emphasized that the neutral definition does not beg such questions as whether presentation of a US following a CR constitutes the same reinforcing operation as presentation of a US on a trial when the CR did not occur; these two procedures represent instances of the same reinforcing operation if and only if they produce the same change in the probability of evocation of the CR by the CS. However, it seems strategic to avoid issues of this sort when, as in the present investigation, we are concerned with the nature of the changes in response tendencies during learning rather than with the conditions giving rise to these changes. Consequently, in the experiments to be reported, we have attempted so far as possible to avoid the customary confounding of reinforcement with antecedent response. In paired-associate situations, for example, we have deviated from the usual anticipation procedure by separating the reinforcement (paired-presentation of stimulus and response members of an item) from the test for learning (presentation of the stimulus member alone) so that an item may receive more than one reinforcement before the first test trial or may receive repeated test trials without intervening reinforcement.

For purposes of measuring retention, it would be ideal if one could give test trials on which no learning at all occurred. Indeed, so long as "learning" is conceived solely in terms of the definition given above (increase in probability of the "correct" response to a given stimulus), this goal is not too difficult to approximate. It seems intuitively clear, and can be demonstrated empirically (Estes, Hopkins, & Crothers, 1960), that no systematic increase in probability of correct responses to, say, paired-associate items will occur over a series of trials on which the stimulus members are presented alone and the *S*'s responses receive no reward or informational feedback from the experimenter. We cannot, however, rule out the possibility that on these trials there might be learning in the sense of an increase in probability of whatever responses, correct or incorrect, actually occur. In fact, there is evidence that such learning does occur, but at a relatively low rate compared to the learning that occurs on reinforced trials (Estes et al., 1960). Consequently, in the analyses to follow, I shall assume that unreinforced trials can be treated, without serious error, simply as "neutral" test trials when primary interest is in measuring the effects of preceding reinforced trials.

UNITARY ASSOCIATION VS. HABIT HIERARCHY

If the stimulus-response relation established by reinforcement were a unitary, all-or-none connection of some sort, then the learning of a new response to a stimulus would automatically displace a previously associated response. Contrariwise, in every variant of the concept of associative strength, it is assumed that a number of different responses may simultaneously be associated with the same stimulus, the relative strengths of association depending primarily upon preceding frequencies of reinforcement of the different responses in the presence of the stimulus. This latter assumption has been embodied in Hull's "habit family hierarchy."

To spell out the question at issue in experimental terms, let us suppose that first Response A and subsequently Response B have been reinforced in the presence of a given stimulus (or stimulus complex—the nature of the stimulation does not matter so long as it is the same on each trial) for each member of a group of *S*s and that a test trial now reveals an observed probability (relative frequency) of .5 for each response. Does this mean that for each *S* both Response A and Response B now have approximately equal habit strengths and therefore equal probabilities of evocation by the given stimulus? Such is the interpretation required by the conception of associative strength (habit strength, excitatory potential, proportion of conditioned elements, or whatever) as well as by any of the contemporary models formulated in terms of continuously variable response probabilities for individual *S*s. Stated in these concrete terms, however, it does not seem that the assumption need depend for support only on extremely indirect evidence, as has hitherto been the case. To test these implications of the strength concept quite directly, we apparently need only reinforce two different responses to the same stimulus for an individual *S*, and then, by means of a series of unreinforced test trials, give the two responses an opportunity to exhibit their relative strengths—if such exist. The predictions under test would be straightforward. On the hypothesis of associative strength, or habit family hierarchy, we expect individual

*S*s to shift back and forth between the two previously reinforced responses over a series of tests. On the hypothesis of a unitary, all-or-none association between stimulus and response, we expect that an individual *S* who makes Response A, say, on the first test trial with a given stimulus will not shift to B on subsequent tests.

A paired-associate learning situation offered a number of convenient features for an empirical realization of the hypothetical experiment just described. The stimuli and reinforcements are readily controllable; and, by embedding a stimulus to which two different responses are to be reinforced in a conventional list, one can make *S*'s task appear little different from familiar learning situations. For each of 20 *S*s, the procedures were replicated with six eight-item lists, all made up on the same principles.[2] The stimuli were all consonant syllables and the responses one-syllable words. Half of the items in each list had single correct responses; examples are:

STIMULUS	RESPONSE
HTX	wish
JFR	sped

These were simply "ballast" and will not be considered further. The other half of the items had two correct responses for each stimulus; examples are:

STIMULUS	RESPONSES
DGR	thaw, weep
BCG	pink, rule

The two correct responses to each stimulus were reinforced equally often for each *S*. By "reinforcement," in this context, we mean simply a paired presentation of stimulus and correct response to *S*; by "test," a presentation of the stimulus member of an item alone. In half of the replications, there were exactly two reinforcements preceding the test trials, one on each response; in the remaining replications, there were four reinforcements, two on each response, prior to the test trials. Following the reinforcements, a series of unreinforced test trials was given, the order of the stimuli being randomized anew for each test.

The critical data for our purposes are the proportions of cases in which the response given on the first test trial to a given stimulus was repeated on the second test. According to an interpretation in terms of associative strength, or habit family hierarchy, repetitions (AA or BB) and shifts (AB or BA) should have occurred with roughly equal frequencies.[3] According to the notion of a unitary association, AA and BB should have each occurred about half of the time and the shifts AB and BA not at all. The results were as follows: Following two reinforcements, there were 85% repetitions and 15% shifts on the first two test trials, the percentages being based on all items that had correct responses on both tests. Following four reinforcements, these values were 89% for repetitions and only 11% for shifts. On later pairs of tests, the frequencies of AB and BA shifts were even smaller. Thus the results do not offer very impressive support for the assumption that a habit hierarchy exists in the individual *S* following the reinforcement of two different responses to the same stimulus. On the other hand, the data appear quite harmonious with an assumption of unitary association, since the proportions of AB and BA shifts appear small enough to be attributable to minor uncontrolled factors such, for example, as fluctuations in context (background stimulation, stimulus traces from preceding items) from one trial to the next.[4]

---

[2] This experiment was conducted at Indiana University with the assistance of E. J. Crothers.

[3] The response given on the first test trial was the more recently reinforced response in approximately 60% of cases. Thus more instances of AA than BB would be expected; nevertheless the proportions of repetitions and shifts expected on the associative strength hypothesis are nearly equal (.52:.48).

Although the present experiment failed to produce the anticipated direct support for the concept of associative strength, a "strength theorist," particularly one working within the framework of Hull's system, may not be too disturbed. With the benefit of an ad hoc assumption concerning the range of "behavioral oscillation," the model of Hull and Spence can be made to yield the prediction that, of two responses reinforced to the same stimulus, one or the other will dominate on test trials, thereby preventing shifts from one to the other on the part of individual *S*s. In order to obtain a more decisive test of the strength conception, we evidently require an even simpler experiment in which only one response is ever reinforced.

With the hope of forestalling an indefinite regress through a series of progressively more refined experiments, suppose we ask what is the minimum set of operations and observations actually needed in order to demonstrate learning. Normally there must be a pretest in order to determine the initial probability of the to-be-learned behavior in the test situation; in practice the experimenter often has a priori information about initial response probabilities which makes the pretest dispensable. There must be a presentation of some reinforcing operation, and afterward a test to assess the change in performance produced by the reinforcement. If the function of response occurrences and nonoccurrences is to be determined, there will have to be a second test trial. And there we have it. Controlled comparisons relative to effects of the principal events occurring during an acquisition series can, in principle, be accomplished in an experiment running to about a trial and a half. By usual standards, this constitutes what can only be called a "miniature experiment." However, miniature experiments appeared to be what the tactical situation called for, and therefore miniature experiments are what we set out to run.

CONCEPTIONS OF THE ACQUISITION PROCESS: ASSOCIATIVE STRENGTH VS. ALL-OR-NONE MODELS

In the first of these experiments,[5] we used a paired-associate situation with consonant syllables as stimuli and numbers as responses. Forty-eight *S*s were run with an eight-item list, yielding 384 observations on the first test trial. The principal portion of the experiment consisted simply in presenting each *S* once with each stimulus-response pair and then testing with each stimulus alone (in a new random order). Before pro-

---

[4] It should be noted that these results cast doubt on the concept of habit hierarchy only as applied to response probabilities in the presence of a stimulus situation that has been manipulated as a unit over the series of reinforced and test trials. M. S. Schoeffler (personal communication) and, in another unpublished study, Crothers and I have obtained positive evidence for the existence of a habit hierarchy associated with a compound stimulus, components of which have been separately correlated with reinforcements during the training trials. The results with compounding provide a control for the present experiment. If the very high proportions of repetitions reported above were attributable to learning that occurred on the first test trial, or to some nonassociative variable, then a similar excess of repetitions should have occurred over successive tests with stimulus compounds; however, no such excess was observed either in Schoeffler's study of compounding or in ours.

[5] This experiment was conducted at Indiana University with the assistance of B. L. Hopkins; for a full report of the method and results see Estes et al. (1960).

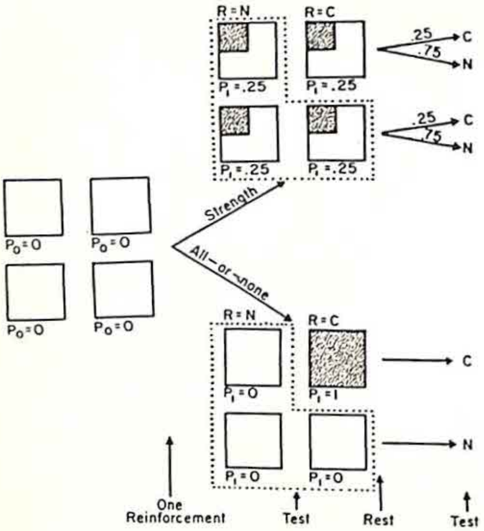


Fig. 1. Schema representing effects of a single reinforcement according to incremental (upper branch) vs. all-or-none (lower branch) theories. Squares represent *S*s, with the proportion of darkened area in each indicating the probability of the correct response (C) for the given individual.

ceeding to the results, let us examine the outcome expected on the basis of the notion of learning as a change in associative strength. In Figure 1 the situation is schematized in terms of a single item. The four squares at the left represent four hypothetical *S*s, the emptiness of the squares indicating that all start the experiment with zero probabilities of making the correct response. Now we give a single reinforcement (paired presentation of the stimulus and correct response), the result of which is to raise the probability of the correct response (C) to, say, .25. The upper arrow leads to the theoretical state of affairs after this reinforcement, according to an interpretation based on the conception of associative strength. The strength of the association is increased for all of the *S*s; and, neglecting for the moment possible individual differences, the probability of the correct response is now .25 for each individual, the one at the upper right who happened to make a correct response on the test, and the three who did not.

Suppose now that the interpretation based on the concept of strength were completely wrong and that stimulus-response associations really formed on an all-or-none basis. Then the state of affairs after the reinforcement should be as shown in the lower part of the figure. Again the probability of a correct response increases from zero to .25, but the .25 now refers only to the group, not to any individual *S*. One *S* has formed the association (darkened square), and three have been unaffected by the reinforcement (empty squares).

To distinguish empirically between these two logically possible outcomes, we need only add the remaining half-trial to our trial-and-a-half, i.e., give another test without intervening reinforcement. Now, if the upper branch of the diagram is essentially correct, all *S*s should have equal probabilities of making the correct response on the second test trial, regardless of what they did on the first test. But if the lower branch is correct, correct responses on the second test should come only from *S*s who made correct responses on the first test. None should come from *S*s who made incorrect responses (N) on the first test, for these *S*s would not have profited at all from the learning trial.

There might be some attenuation of the expected proportions of correct responses on Test 2 by *S*s making correct responses on Test 1 if there is any forgetting in this situation, but the proportions of correct following incorrect provide a critical comparison. If the all-or-none view is correct, then this proportion should be zero, or at least no greater than could be achieved by sheer guessing. But if any version of the strength conception is correct, then the proportion of correct following incorrect responses should be greater than chance. In order to make the outcomes that can be tolerated by the two interpretations sharply different, we need only choose our experimental materials and conditions so that the overall proportions correct on both first and second tests are well above chance. It can be seen in Figure 2 that this has been achieved, for approximately 50% of the items were correct on the first test and nearly 40% on the second. Considering the critical lower branch of the diagram, leading from an incorrect response on the first test to a correct response on the second, we see that the results lean strongly in the direction prescribed by an all-or-none conception, for the 9% of correct following incorrect responses is less than the 12½% that could be achieved even by rather unintelligent guessing with an eight-item list if the reinforcement had no effect at all on these items. The difference between this value and the 71% of correct following correct responses is so large that a statistical test would be an empty formality.

A possible defense that might be advanced by a "strength theorist" is the hypothesis that the 51% of cases with incorrect responses on the first test simply represent preponderantly slower learners or more difficult items than the 49% of cases with correct responses. If so, then a control condition in which a second reinforcement is given between the first and second tests should yield a percentage of correct responses on Test 2 following incorrect on Test 1 that is much smaller than the percentage correct on Test 1. This control was run (with the same 48 *S*s but different items), and the result is shown in Figure 3. The effect of the first reinforcement on the full set of *S*s and items was to raise the probability of a correct response from near zero to .40; the effect of the second reinforcement on cases having incorrect responses on the first test was to raise the probability of a correct response from near zero to .46. Thus there seems to be no support forthcoming for the hypothesis of a large difference in learning rate between cases which did and cases which did not have correct responses on the first test.

Although it would be nice to claim credit for rare prescience in predicting the outcome of this little experiment, the fact is that the result came as a distinct jar to my preconceptions. In designing the study, our idea was not to undermine the strongly entrenched concept of associative strength, but to support it by showing that the results of Rock's experiments, apparently calling for an all-or-none interpretation, must be attributed to some artifact concealed in his ingenious but somewhat complex procedures. Thus when Hopkins and I examined the data from our initial group of 24 *S*s and found the pattern shown in Figures 2 and 3, our first reaction was to replicate the whole thing with another group. But when the two replications turned out to agree in every essential respect, we were left

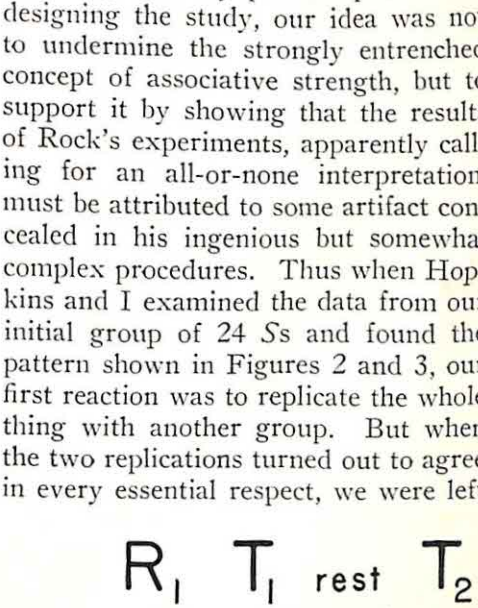


FIG. 2. Results of miniature experiment on acquisition of paired-associates. Empirical values are proportions of instance in which correct (C) and incorrect (N) responses on first test trial after a single reinforcement (paired presentation of stimulus and response members) were followed by C and N responses on a second test trial.

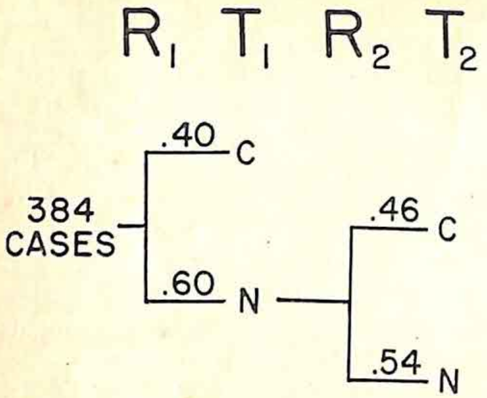


FIG. 3. Proportion correct on test after a second reinforcement for cases not having correct responses on first test compared with proportion correct on first test for the full set of *S*s and items. The *S*s and situation are the same as those represented in Figure 2.

with no obvious course but to begin digesting an unanticipated and not entirely palatable conclusion. The most cleanly controlled comparisons we had managed to devise yielded no evidence that repeated reinforcements in this situation have any function other than to give repeated opportunities for the discontinuous formation of a learned association between observed stimuli and responses.

Still, it is well known that theoretical doctrines do not yield readily to negative evidence. One whose theories are based on a concept of strength will lose little ground if he can make a stand on the claim that all-or-none acquisition is simply a peculiarity of the paired-associate experiment and not characteristic of human learning in general. To evaluate this possible defense of the strength concept, we clearly shall have to turn to some different situation that is quite different in the response mechanism and reinforcing operations from paired-associate learning. Eyelid conditioning meets these specifications, and it is convenient for our purposes since a colleague, I. Gormezano, has kindly made available his data from an intensive period of data collecting in the Wisconsin conditioning laboratory. Gormezano trained a sufficiently large group, approximately 170 *S*s, under identical conditions so that the first few acquisition trials can be treated as one of our miniature experiments and analyzed in much the same way as the paired-associate study.

The situation obtaining over the first couple of trials is schematized in Figure 4. In the diagram, T_1 is the first CS presentation, prior to the first reinforcement, and we shall consider only *S*s who made no CR on this test. Thus the initial probability of a CR is taken to be zero. Suppose now that the effect of the first reinforcement is to raise the probability of a CR to .25. According to the strength conception, shown in the upper panel, each *S* has his strength of conditioning increased by the same amount by this reinforcement and now has probability .25 of making a CR. Then the second reinforcement increases the conditioned strength for each *S* again; and, regardless of whether or not a particular *S* happened to make a CR on T_2, he now

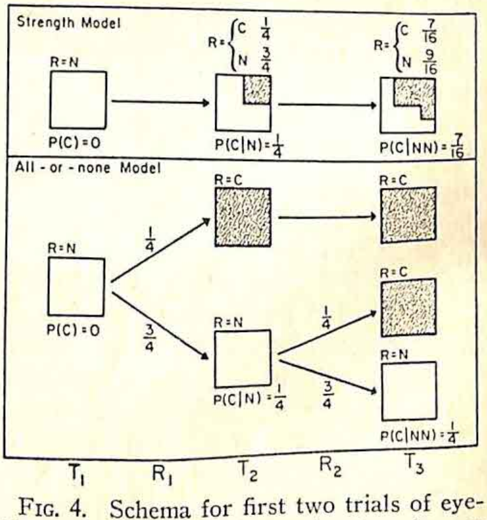


FIG. 4. Schema for first two trials of eyelid conditioning experiment showing changes in CR probability (proportion of darkened area in squares representing *S*s) prescribed by incremental vs. all-or-none theories.

has a higher probability ($\frac{7}{16}$ if we apply the linear function mentioned earlier).[6] According to an all-or-none conception, the situation after the first reinforcement, shown in the lower panel, is that for ¼ of the *S*s the CR has become associated with the CS and for the remaining ¾ of the *S*s no conditioning has occurred. The effect of the second reinforcement is to give the unconditioned *S*s another chance, and ¼ of these now become conditioned. The differential prediction, then, concerns the probability of a CR on the third test for *S*s who made no CR on the second test (and similarly the probability of a CR on the fourth test for *S*s who made none on any previous test, and so on). The strength conception requires this conditional probability to increase, whereas the all-or-none conception requires it to remain constant. The test seems quite sharp, for even with allowance for variation in conditioning rates among *S*s, a model which assumes that associative strength increases with reinforcements cannot stand constancy of this probability unless its assumptions are so restricted that it reduces to an all-or-none model.

The pertinent results of Gormezano's study are shown in Figure 5, carried through the first four trials, beyond which the number of cases begins to drop off too much for comfort. Inspecting the sequence of probabilities of CRs after 1, 2, or 3 consecutive NCRs—.13, .15, .14—we find the hypothesis of constancy appearing rather more attractive than the progressively increasing trend required by the strength interpretation. (According to a linear model, for example, the value .15 for a CR after an NCR should have been .24, and the .14 for a CR after two NCRs should have been .34.)

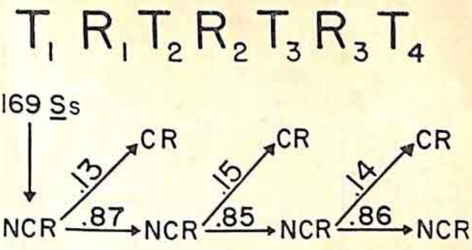


FIG. 5. Trial-by-trial acquisition data from Gormezano's study of eyelid conditioning. Values of particular interest are the proportions of CRs following 1, 2, or 3 consecutive non-CRs.

The consistency of these conditioning data with those of the paired-associate situation is almost too good to be true. In psychology we are not used to having quantitative tests of alternative theoretical notions yield such apparently decisive outcomes. Consequently, and considering the importance of the theoretical issue, perhaps we will not yet be accused of beating a dead hypothesis if we look for one more test with experimental arrangements differing from both of those preceding. We would like a situation similar to paired-associate learning in that unreinforced test trials can readily be given without disturbing the learners but one which eliminates the possibility of achieving substantial proportions of correct responses by guessing. A situation which meets these desiderata is the free verbal recall experiment used by Bruner, Miller, and Zimmerman (1955). For our present purposes the minimal experiment will consist of a single reinforcement followed by two recall tests. The reinforcement involves merely the experimenter's reading a list of words aloud to *S*. On a recall test, *S* is asked to write down as many words as he can

---

[6] For this example, the parameter θ in the function $\Delta p = \theta(1 - p)$ is equal to ¼ and after the first experiment p is also equal to ¼. Therefore we have

$$\Delta p = \tfrac{1}{4}(1 - \tfrac{1}{4}) = \tfrac{3}{16}$$

and for the new probability after the second reinforcement,

$$p + \Delta p = \tfrac{1}{4} + \tfrac{3}{16} = \tfrac{7}{16}$$

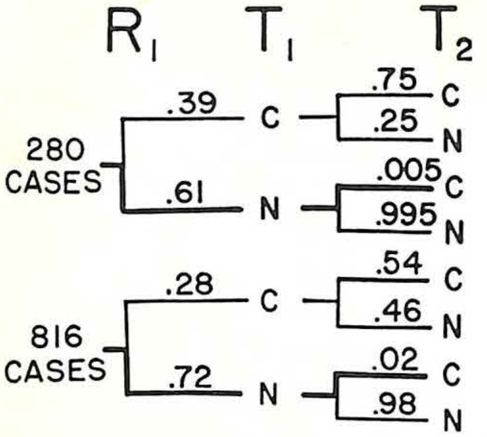


FIG. 6. Results of two miniature experiments on free verbal recall, showing near-zero proportions of correct responses on a second test trial for cases which did not have correct responses on the first test after a single reinforcement.

remember (in any order). Then after an interval during which no additional reinforcement is given, *S* is (unexpectedly) tested again.

Results of two experiments[7] of this sort are shown in Figure 6. The upper tree represents an experiment with 35 *S*s, each given a list of eight words at R_1. On the first test, T_1, 61% of the 280 opportunities for correct responses (C) yielded either incorrect responses or omissions (N), and of these less than 1% were followed by correct responses on the second test, T_2. In a replication conducted with some minor variations in procedure, 102 *S*s were presented with eight words each on the reinforced trial. This time (lower tree in Figure 6) there were 72% N responses on the first test and less than 2% of these were followed by C responses on the second test. Clearly, if a word is not given correctly on the first test by a particular *S*, the chances are virtually *nil* that it will be correct on a second test.

This result does not, of course, *prove* that reinforcement has exerted no strengthening effect on the associations in the cases when the correct response failed to occur on the first test. But one whose theory requires him to assume that such strengthenings occur has a taxing assignment in producing a case for the existence of factors or processes which appear in just sufficient force to cancel out the hypothesized increments in response strength under each set of experimental procedures we have examined. Explanations depending on such factors as individual differences are not very prepossessing in the light of control comparisons of the type exhibited in Figure 3 (a similar control, with a similar result, was used for the free-verbal-recall situation). One might appeal to the effects of learning which occurs on the first test trial itself, arguing that an incorrect response which occurs on the first test receives a large increment in associative strength (from sheer contiguity or perhaps from some unspecified source of reinforcement) and therefore recurs with high probability on the next test. One important difficulty with this hypothesis is that the data do not support it. In the paired-associate study cited above, for example, the observed relative frequency with which an incorrect response occurring on the first test was repeated on the second test was only .24. Interpretations which preserve the incremental conception of associative learning should certainly be sought with all vigor; at the time of writing, however, none has come to my attention that seems at all plausible.

RETENTION AS A FUNCTION OF REINFORCEMENT AND INTERSPERSED TEST TRIALS

The story does not end here. Even if one is ready to grant that associations are made and broken in an all-or-none fashion, this concession, although by no means a minor one, does not exhaust the resources of the strength concept. It is possible that after an association has once formed, associative strength in the sense of resistance to forgetting continues to grow as a function of reinforcements. In fact the experimental situations we have used appear well suited to demonstrate such an effect.

---

[7] These experiments, conducted at Indiana University with the assistance of Judith Crooks, will be reported in detail elsewhere.

It will not have escaped notice that although the three miniature experiments yielded sharply negative results relative to the conception of learning as a gradual strengthening of associations by reinforcement, they agree only in part with the usual notion of all-or-none learning (as, for example, defined by Voeks, 1955). The formation of learned associations appears to be discontinuous rather than gradual, but once a correct response (or conditioned response) has occurred to a given stimulus, it does not appear with certainty when the stimulus recurs after a rest interval. In fact, "forgetting" as measured by the proportion of shifts from a correct response on the first test to noncorrect on the second test ranges from about 20% in the paired-associate situation to about 50% in free verbal recall. For the eyelid conditioning study, we did not report these proportions since the numbers of observations on which they were based were rather small; but for what it is worth, we might add that the proportions of CR to NCR shifts following the first three tests, despite the intervening reinforcement, fall well within the range given above for the other two experiments.

The minimum set of operations and observations required to test for a dependence of "resistance to forgetting" upon number of reinforcements is exhibited in Table 1. Following either one or two reinforcements, a sequence of two unreinforced test trials is given; and retention is measured in terms of the proportion of instances in which a correct response to a given stimulus on the first of these tests is repeated on the second test. Results of two such experiments are shown in the table.

The left-hand column of proportions represents data from a study conducted by the writer with the assistance of E. J. Crothers (Estes et al., 1960). Twenty *S*s were each run on six eight-item paired-associate lists under each condition (number of reinforcements). The stimuli were nonsense syllables, and the responses were numbers in half the lists and familiar words in the other half. The right-hand column presents data from the free verbal recall experiment previously cited (102 *S*s each tested with a list of eight items under each condition). The pattern of test proportions needs little comment. In each case the difference called for by the traditional conception of associative strength fails to appear.

This outcome is a little hard to swallow. It is well known that retention increases with overlearning. If the additional reinforcements given during the overlearning period do not produce the increased resistance to forgetting, then what does? We may obtain a clue as to the answer by introducing what might a priori seem to be very slight change in the design of our minimal experiment on retention. This variable is illustrated in Table 2. In the free

TABLE 1

DESIGN AND RESULTS OF MINIMAL EXPERIMENT ON RETENTION AS A FUNCTION OF NUMBER OF REINFORCEMENTS

Procedure	Amount of Retention P-A	 FVR
R T_1 T_2	.90	.54
R R T_1 T_2	.89	.52

TABLE 2

DESIGN AND RESULTS OF MINIMAL EXPERIMENT ON RETENTION AS A FUNCTION OF A TEST TRIAL INTERPOSED BETWEEN REINFORCEMENTS

Procedure	Amount of Retention
R R T_1 T_2	.52
R T_1 R T_2 T_3	.78

verbal recall experiment cited previously, we included, in addition to the condition shown in the upper row (which is the same as in Table 1), a condition with the same number of reinforcements but with an additional interspersed test trial. The idea was to give an opportunity for correct responses learned on the first reinforced trial to become conditioned to cues which are present only on test trials and not on training trials—in other words to spread the effect of the first reinforcement over more stimuli. Retention is measured in terms of the proportion of correct responses repeated from T_1 to T_2 in the upper row and from T_2 to T_3 in the lower row. In contrast to the lack of effect on retention produced by increasing the number of reinforcements, note the large effect of the added test trial. Two reinforcements without an intervening test yield only 52% retention, whereas two reinforcements with an intervening test trial yield 78% retention from the first to the second test following the second reinforcement.

Although the principle that retention increases with number of reinforcements is exceedingly well established in the lore of human learning, we must face the possibility that this empirical relation, like the classical acquisition curve, is an artifact of the confoundings inherent in the usual experimental paradigm.

CONCLUDING REMARKS

To recapitulate the box score: we have conducted a series of highly simplified experiments especially designed to provide relatively direct evidence for the widely accepted interpretation of learning and retention in terms of increments and decrements in associative strength. We noted that in virtually all contemporary learning theories, the concept of strength is assumed to have at least three different empirical manifestations—the habit hierarchy, the growth of associative strength in the sense of response probability as a function of number of reinforcements, and the increase in resistance to forgetting of once established associations with additional reinforcement. We tested for all three of these effects by means of the simplest controlled comparisons we could arrange, doing our utmost to eliminate the confoundings and the layers of statistical processing that shield the behavioral changes occurring on individual learning trials completely from view in conventional experimental designs. And under these presumably favorable circumstances, all three effects mysteriously evaporated, leaving a picture of unitary associations the learning and unlearning of which proceed on an essentially all-or-none basis.[8]

---

[8] Several readers of a prepublication draft of this article have raised the question whether alternative measures, e.g., response latencies or recognition scores, might yield evidence of learning in cases where the probabilities of conditioned responses or of correct recalls do not. Regardless of the answer, it is important to note that information about concomitant changes in other variables would have no logical bearing upon conclusions pertaining to the one actually chosen for analysis in a given experiment. To determine whether the behavior of recalling the correct response to a paired-associate item is learned in all-or-none fashion, we required an analysis according to the paradigm of Figure 2 with recall score as the dependent variable. To determine whether

What is the import of these results for the question raised earlier concerning the strategic soundness of the general methods on which investigators of learning chiefly rely? The concept we have examined in detail epitomizes the intervening-variable paradigm for theory construction that has been popularized by Tolman and Hull, and their followers, to the point of dominating contemporary learning theory. The general technique is to postulate a hypothetical state or entity which is held to intervene in some sense between observed stimulus and response variables. Predictions derived from theories built around such constructs are checked against data from standard learning experiments. Thus the adequacy of the theories depends in turn on the adequacy with which essential aspects of learning are captured by standard experimental designs.

I have emphasized the term "standard" for despite the fact, decried by many critics, that psychologists in this field have resisted all admonitions to standardize their tasks, situations, and procedures, they have more than compensated for this lack by the degree to which they have standardized their experimental designs. Scanning the reference list of, say, Hilgard's *Theories of Learning* (1956),[9] one can find no more than a meager handful of studies which are not molded into a paradigm that might be termed the groups-by-trials design. The master blueprint requires the investigator to average some type of performance score, e.g., frequency of conditioned responses or of correct responses, over a series of trials for groups of *S*s—the groups being differentiated on the basis of conditions obtaining over the series. Going back a few years in the experimental journals, one finds the customary output of this experimental paradigm to be a set of mean performance curves for experimental and control groups; more recently the output is typically an array of analysis of variance tables, perhaps supplemented by some mean criterion scores. These observations in themselves do not constitute a criticism. No one would gainsay that the groups-by-trials design and its associated statistical techniques are useful tools for assaying the effects of various procedures and conditions upon performance. It is easy, however, to overlook the fact that the groups-by-trials design yields only information about relationships and trends which hold on the average over groups of *S*s and series of trials. No accumulation of experiments, however large, all conducted and all analyzed in accord with this same general method can provide a sufficient empirical check on concepts and assumptions that refer to processes or events occurring in the individual learner. The findings we have considered in this paper suggest that in point of fact some of the most firmly entrenched concepts and principles of learning theory may be in a sense artifacts of a conventionalized methodology.

The laboratory investigator of learn-

---

correct recognition of a previously viewed item is learned on an all-or-none basis, we would need a similar analysis with recognition score as the dependent variable. To determine whether the behavioral change associated with a decrease in latency is learned on an all-or-none basis, we would need a similar analysis with some criterion of change in latency as the dependent variable, and so on.

[9] The harvest is even scantier if one consults the more monographic works, e.g., Tolman (1932), Hull (1943), or Spence (1955), associated with systems of intervening-variables in behavior theory; or, for that matter, my own theoretical writings (Estes, 1959). In the eclectic treatise of McGeoch and Irion (1952), deviant experimental designs appear in references with the earliest dates but disappear as one comes down to the contemporary literature.

ing is used to going his own way more or less oblivious to the rattle of criticisms from textbook writers, educators, and other "outsiders." Traditionally the criticisms have always been to the same effect—that the "learning experimentalist" should push on from the worn out fields of conditioning and simple verbal learning, where there are really no unsolved problems of any importance remaining, and devote himself to richer, more complex, experiments that come closer to learning situations of real life (as contrasted with the unreal life found in the laboratory). Here we break with tradition, for the criticism generated by our work within the field suggests that progress toward a satisfactory theory of learning requires, not more complex, but simpler experiments. The conventional experiment, far from being oversimplified, represents such a complex and intricate confounding of stimulus and response variables over trials that once it has been done and reported in conventional form, no amount of study of the analysis of variance tables and Vincentized performance curves can disclose the effects exerted by specific causal variables on individual *S*s upon particular occasions. Concepts depending solely upon the conventional experiment for support may turn out to belong, not to a psychology of learning, but only to a psychology of the criterion score, the mean performance curve, and the groups-by-trials design.

In concluding, I would like to indicate that I do not mean to offer the experimental findings reported in this paper as a crucial test of incremental vs. all-or-none theory. What does seem clear with respect to the former is that the kind of evidence heretofore adduced in support of incremental theories is inadequate to distinguish them from alternative conceptions. More penetrating experimental analyses are required. We have made a start in this direction, and from the early returns it appears that no extant theory of the incremental type can handle the pattern of results that is emerging. The temptation is great to indulge now in a bit of speculation as to whether the answer to this situation will prove to be a remodeling of one of the familiar incremental theories, or perhaps a quite different theory based on all-or-none assumptions—but this is a step I do not intend to take.

Since the earliest days of associationism, overdependence upon speculation and circuitous inference has impeded the interplay of theory with experiment. While I would not for a moment depreciate the role of imagination in science, I suspect that it will begin to serve us effectively in learning theory only as we begin to accumulate reliable determinations of the effects of single variables upon single learning trials in individual organisms. If by continual simplification of our experimental analyses and refinement of our mensurational procedures we can achieve these determinations, we may find that the long sought laws of association may be not merely "instigated," or even "suggested," but literally dictated in form by empirical data.

REFERENCES

BRUNER, J. S., MILLER, G. A., & ZIMMERMAN, C. Discriminative skill and discriminative matching in perceptual recognition. *J. exp. Psychol.*, 1955, **49**, 187–192.

BUSH, R. R., & MOSTELLER, F. *Stochastic models for learning.* New York: Wiley, 1955.

DEESE, J. *The psychology of learning.* (Rev. ed.) New York: McGraw-Hill, 1958.

DUFORT, R. H., & KIMBLE, G. A. Ready signals and the effect of interpolated UCS presentations in eyelid conditioning. *J. exp. Psychol.*, 1958, **56**, 1–7.

ESTES, W. K. Learning. *Annu. Rev. Psychol.*, 1956, **7**, 1–38.

ESTES, W. K. The statistical approach to learning theory. In S. Koch (Ed.), *Psychology: A study of a science.* Vol. 2. New York: McGraw-Hill, 1959. Pp. 380–491.

ESTES, W. K., HOPKINS, B. L., & CROTHERS, E. J. All-or-none and conservation effects in the learning and retention of paired associates. *J. exp. Psychol.,* 1960, in press.

GOODRICH, K. P., ROSS, L. E., & WAGNER, A. R. Performance in eyelid conditioning following interpolated presentations of the UCS. *J. exp. Psychol.,* 1957, **53**, 214–17.

HILGARD, E. R. *Theories of learning.* (Rev. ed.) New York: Appleton-Century-Crofts, 1956.

HULL, C. L. *Principles of behavior.* New York: Appleton-Century, 1943.

KENDLER, H. H. Learning. *Annu. Rev. Psychol.,* 1959, **10**, 43–88.

KIMBLE, G. A., & DUFORT, R. H. The associative factor in eyelid conditioning. *J. exp. Psychol.,* 1956, **52**, 386–91.

KIMBLE, G. A., MANN, L. I., & DUFORT, R. H. Classical and instrumental eyelid conditioning. *J. exp. Psychol.,* 1955, **49**, 407–17.

MCGEOCH, J. A., & IRION, A. L. *The psychology of human learning.* (Rev. ed.) New York: Longmans, 1952.

RESTLE, F. A survey and classification of learning models. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory.* Stanford, Calif.: Stanford Univer. Press, 1959.

ROCK, I. The role of repetition in associative learning. *Amer. J. Psychol.,* 1957, **70**, 186–193.

SPENCE, K. A. *Behavior theory and conditioning.* New Haven: Yale Univer. Press, 1955.

SPENCE, K. A. Behavior theory and selective learning. In M. R. Jones (Ed.), *Nebraska symposium on motivation.* Lincoln: Univer. Nebraska Press, 1958.

TOLMAN, E. C. *Purposive behavior in animals and men.* New York: Appleton-Century, 1932.

VOEKS, V. W. Gradual strengthening of S-R connections or increasing number of S-R connections. *J. Psychol.,* 1955, **39**, 289–99.

(Received October 26, 1959)

Psychological Review
1960, Vol. 67, No. 4, 224–228

ON PROBLEMS OF CONDITIONING DISCRIMINATED LEVER–PRESS AVOIDANCE RESPONSES[1]

DONALD R. MEYER, CHUNGSOO CHO, AND ANN F. WESEMANN
Ohio State University[2]

The learning of avoidance, as Solomon and Brush (1956) have remarked, is something that we all take for granted. But, and they have put it particularly well, the

> failure of *S*s to learn to avoid are *not* rare, but they are less apt to be reported in scientific papers than are the successes. And the conditions surrounding failures to learn to avoid are as instructive to the investigator, in many cases, as are the optimum conditions for producing the avoidance phenomenon.

It is in this spirit that we offer certain data concerning failures to avoid. These have been collected in both formal and informal studies that have now occupied the better part of some two years. All have been conducted with rats, and all within the format provided by one or another leading brand of commercial lever-pressing apparatus. Some slight differences have been observed between the different brands, but we can conclude at the outset that the problems herein described are not unique to any variant. Rather, the difficulties seem to be inherent in the lever-pressing situation, and not in different constellations as provided by the different fabricators.

THE BASIC PROBLEM

We must emphasize that such equipment *can* be used in training rats to make stimulus-discriminated lever-press avoidance responses. Our concern is wholly with the rate at which such learning is established. We had first obtained this apparatus for some studies of the functions of the visual cortex; these investigations had been planned as extensions of the work of Meyer, Isaac, and Maher (1958). The bar-press situation was selected because we wanted to examine two responses that would be the same except for the incentive under which their learning was established. Further limitations thus imposed were two: that light be the warning stimulus, and that the habit be learnable within a single day. These did not, at first, appear to be unreasonable requirements. The primitive device that we had used before had worked well enough to give us learning in very rarely more than 100 light-shock presentations. It was our presumption that the lever-pressing habit could not be established quite so quickly, but we readily accepted this because of the convenience that the new equipment afforded. We were not, however, prepared in any way for our first empirical encounter, during which a rat failed to show reliable avoidance after *several thousand trials*.

A few such experiences led us to conclude that we were doing something very wrong; it did not seem possible that lever-press conditioning could be as bad as this for the purpose. However, the first of our inquiries to persons familiar with the bar-press situation did not prove to be encouraging; we learned, for example, that it often takes as much as 40 hours to establish stable levels of performance.

---

[1] This research was supported in part by a grant from the National Institutes of Mental Health.

[2] From the Laboratory of Comparative and Physiological Psychology.

Somewhat more salubrious estimates were nearer to 10, but it seemed that not all rats could be counted on even then. This perhaps should not have shocked us, but it did—for we were remembering with sorrow that Mote (1940) had trained his average rat to turn a wheel within 50 light-shock presentations. The figure of merit for the lever-press response is, by this comparison, so poor that only in an automated era could it possibly survive as an indicant of choice.

In retrospect, so far as our brain work was concerned, our course of action at this point should have been to build a Brogden-Culler apparatus and give up lever pressing forthwith. But it is most difficult to turn one's back to a gleaming, thousand-dollar box; then too, we were still convinced that *some* combination of factors would produce efficient learning. Therefore we proceeded to assess the most obvious of these in an informal manner; bar weights, shock strengths, and light intensities were among the first considerations. Throughout these experiments, the general conditions were all very much the same: the warning stimulus consisted of a sharp illumination of the box; the box was darkened when the bar was pressed, and this response terminated shock on trials when the shock was not avoided.

There was never any problem whatsoever in training rats to make escape responses: indeed, after relatively few shaping trials—rarely more than 30 or 40—the latencies on most trials were so short that it could be doubted that the shocks received were noxious. Frequently the rat would learn to hover just above the locus of the bar; this position was maintained throughout the presentation of the warning light. When the shock arrived, the rat depressed the bar very quickly, but in most instances it failed to learn to do this during the warning interval. To counter this maneuver of "avoidance by escape" we next arranged a program such that shocks could always be avoided, but that periodically were of a fixed duration rather than escapable. This of all the variables informally considered seemed to make at least some difference.

TWO ILLUSTRATIVE EXPERIMENTS

Two formal studies serve to illustrate just how difficult this problem is. We had hoped, because of certain problems of control, to train at least three rats a day; practically, the most that we could possibly invest in training was a 4-hour session. We were therefore driven to consider massing as a possibility and hence trained three groups of nine rats each with intervals between trials of 20, 40, and 60 seconds, respectively. Each trial consisted of light presentation for a 5-second period; pressing of the bar turned off the light and stopped or prevented the occurrence of a shock of 1.6 milliamperes. The outcome, now so typical of many, was that the animals learned virtually nothing during their first training sessions, and this despite the fact that they had first been shaped to make escape responses and then had been presented with from 240 to 720 light-shock pairings.

Some small evidence of avoidance learning did appear when training was continued for a second daily session. The intermittent nonescape procedure then was instituted, and the rats were trained for three more days. But the final outcome, based upon a total of from 1,200 to 3,600 presentations, was that the 20- and 40-second groups still were not avoiding properly. For these groups, in fact, the median percent avoidance scores were .59 and 1.32. The 60-second group was somewhat better: here at least some animals performed beyond a level of 50% avoidances. Only four, however, did so, and the median performance for the group was mere 6.69%.

At this juncture, we began to speculate that the problem might be in duration of the warning stimulus. This has been manipulated widely in different experimental contexts; Warner (1932), for example, examined a range of from 1 to 30 seconds in a barrier situation. He found that learning was progressively faster as the warning interval decreased, but did not examine the range below 10 seconds in detail. In most relatively recent studies, these have been the intervals of interest, and 5 seconds seemingly is most popular. Mote, on the other hand, presented light for just 2 seconds, and this interval was used successfully by Meyer, Isaac, and Maher. While we could not guess what the optimum would be in lever-pressing situations, it appeared to us that the best place to look was in the range below 5 seconds.

For this experiment, the presentations were given at a rate of one a minute; four groups of rats, 15 in a group, were trained for a total of six hours. There were two daily training sessions, each of which was three hours long; there were thus 180 pairings per day and 360 total pairings. We must say, perhaps, that these are few trials by operant conditioning standards; however, the number is *seven times* as large as that required for learning in a wheel.

The differential treatments were confined to the first training session. Warning stimulus durations prior to shock were, for the four respective groups, 0, .5, 1.5 and 2.5 seconds. During the second session, this was fixed for all at a value of 5 seconds. This was done so that the superiority of one of the differential treatments might be evidenced without the prejudice of changing latency requirements. Here again, as in all the studies we have done, lever-pressing terminated light; if the animal did not avoid, light and shock went off together.

These procedures yielded median performances approximating 20%; variations for the different intervals proved to be no more than 5%. But even these performances could not be taken as evidence for some avoidance learning; associated with them were bar-pressing rates that often reached extremely high values. There were many instances, in fact, of rates above two thousand per hour; had these rates been evenly distributed in time, at least a third of all the subjects would have shown us ultimately perfect *avoidance*. We obtained, instead, a suppression of the rate when the warning stimulus came on; and this suppression was evidenced throughout the conduct of the study.

INTERPRETATION

Suppression, in itself, is not surprising. It is, indeed, the essence of the readily established CER of Hunt and Brady (1951); we expect it and we almost always find it regardless of our test technique. The question, rather, is why such suppression persists in lever-pressing situations, when its existence presumably implies a drive which lever pressing reduces. If light produces "fear," and shock produces pressing, and pressing eliminates the light, how by any concept of avoidance learning can we fail to get avoidance?

We have found but one asylum, and that is in supposing that the offset of light is *not* an effective "fear-reducer." Having done so we can only wonder why we had expected it to be; in ourselves, we know that "fears" persist, once evoked, far beyond their instigators. We do not intend this statement as a glove to Mowrer and Lamoreaux (1951) however; all that they have said is that termination "will have the effect of reducing the rat's fear at least a little and will thus reinforce, at least mildly . . ." the response which brings about the termination. Some reinforcement is undoubtedly produced, and we would merely be inclined to substitute the word "trivially" for "mildly."

But why is the particular form of the response so important? Most experimenters with whom we have discussed this problem have considered it to be a matter of the relative availability of the response to be conditioned. However, it is difficult to argue that responses sometimes made as many as 20,000 times are lacking in availability. We believe, instead, that response topography is of consequence for other reasons and that we can best begin to show that this is so by first reviewing Schoenfeld's (1950) concept of the process of avoidance.

While recognizing that a light, such as we have used, becomes aversive after noxious pairings, Schoenfeld has emphasized the role of proprioception as linking avoidance to escape. In a sense, his view is that the rat, in avoiding, is escaping from its own response to the warning stimulus. Termination of the proprioceptive feedback is a reinforcing operation, as is termination of the now-aversive warning stimulus itself. We believe that this approach is basically correct from the standpoint of its emphasis and that we can extend it to interpret the fact that apparatus can be ordered in efficiency for avoidance learning from wheel to shuttlebox to lever.

Briefly, the argument is this. If the termination of a proprioceptive pattern has a reinforcing function, promptness and completeness will determine how effective this will be. Promptness and completeness will themselves depend upon our getting rid of the responses whose presence at the moment has yielded the proprioceptive pattern to be terminated. Some responses will, of course, disappear upon removal of related stimuli, but others will persist unless inhibited through evocation of antagonists. The freezing-crouching pattern, or conditioned "fear," is a member of this latter group; thus the termination of its proprioceptive pattern will require response competition.

Since the freezing-crouching pattern is a posture, suitable antagonists are few; in specifying these, we think we specify the quickly learned avoidance responses. These responses must, in scope, be commensurate with the pattern of conditioned fear, for when this is only partially the case the inhibition cannot be complete. The lever-press response thus is learned but slowly because it is a poor antagonist; its occurrence can be quite compatible with persistence of the prior pattern. But the situation is obviously different with a jump or running response; these yield us competition, termination —and rapid avoidance learning.

Mowrer (1940), twenty years ago, remarked in passing that he had had no luck when he had tried to establish evidence of good avoidance learning with a "type of pedal-pressing response." With his conclusion we must heartily agree, and we think our tests have been exhaustive. However inconvenient the general implication, operants are *not* arbitrary; in avoidance learning, their selection is perhaps the most important of considerations. Possibly, a standard box can be so modified that suitable responses will appear; we ourselves are trying this in a minor way with a "hit-the-ceiling" pattern. Whether this will work as well as a shuttlebox is something that remains to be seen, but never again will we regard such situations as outmoded and inelegant survivors of an era best forgotten.

REFERENCES

HUNT, H. F., & BRADY, J. V. Some effects of electroconvulsive shock on a conditioned emotional response ("anxiety"). *J. comp. physiol. Psychol.,* 1951, **44**, 88–98.

MEYER, D. R., ISAAC, W., & MAHER, B. The role of stimulation in spontaneous reorganization of visual habits. *J. comp. physiol. Psychol.,* 1958, **51**, 546–548.

MOTE, F. A. Correlations between conditioning and maze learning in the white rat. *J. comp. Psychol.,* 1940, **30**, 197–219.

MOWRER, O. H. Anxiety-reduction and learning. *J. exp. Psychol.,* 1940, **27**, 497–516.

MOWRER, O. H., & LAMOREAUX, R. R. Conditioning and conditionality. *Psychol. Rev.,* 1951, **58**, 196–212.

SCHOENFELD, W. N. An experimental approach to anxiety, escape, and avoidance behavior. In P. J. Hoch & J. Zubin (Eds.), *Anxiety.* New York: Grune & Stratton, 1950. Pp. 70–99.

SOLOMON, R. L., & BRUSH, E. S. Experimentally derived conceptions of anxiety and aversion. In M. R. Jones (Ed.), *Nebraska symposium on motivation, 1956.* Lincoln: Univer. Nebraska Press, 1956. Pp. 212–305.

WARNER, L. H. The association span of the white rat. *J. genet. Psychol.,* 1932, **41**, 57–90.

(Received August 28, 1959)

Psychological Review
1960, Vol. 67, No. 4, 229–242

ON THE TRAINING OF ORIGINALITY[1]

IRVING MALTZMAN
University of California, Los Angeles

There has been a growing research interest in originality during the last decade. Areas of major concern have been the study of the productivity of creative individuals (Dennis, 1958; Lehman, 1953; Taylor, 1956, 1958; Van Zelst & Kerr, 1951), their personality (Barron, 1955; Drevdahl, 1956; Taylor, 1956, 1958), the sociological and cultural factors promoting creativity (Barnett, 1953; Stein, 1953; Taylor, 1956, 1958), and investigations designed to develop reliable paper and pencil tests for the study of individual differences (Guilford, 1950; Springbett, Dark, & Clarke, 1957; Wilson, Guilford, & Christensen, 1953).[2]

Although there is general recognition of the importance of the problem of devising training techniques for facilitating creativity, most of the work in this important area has been anecdotal or merely hortatory (Ghiselin, 1955; Mearns, 1958; Osborn, 1957; Slosson & Downey, 1922).

SCOPE OF THIS PAPER

The purpose of this paper is to review the experimental research that may be relevant to the problem of devising techniques for increasing originality and to indicate the behavioral principles possibly involved in the production of the desired effect.

For purposes of further discussion we shall distinguish between originality and creativity. Originality, or original thinking, as we shall use the term, refers to behavior which occurs relatively infrequently, is uncommon under given conditions, and is relevant to those conditions. Criteria of relevance and uncommonness, of course, must be established for any given situation. This is a problem more readily solvable within the laboratory than without. Creativity, according to the present usage, refers to products of such behavior and the reactions of other members of a society to those products. An invention is a creative product that may have an important effect upon society and is a consequence of original behavior. But considerably more variables enter into the determination of creative works than originality alone. Our distinction implies that an individual may be highly original but not creative. His brilliant theories are never published, or they are not implemented by the necessary research. Perhaps the original research is conducted, but the results are not published or patented. All this may be done, but the work may be overlooked or the implications disregarded by the appropriate community of scholars or society at large. Further common illustrations may be given, indicating that a great many more behavioral and societal variables influence creativity than originality, making the study of originality under simplified laboratory conditions more feasible than that of creativity. Many of these variables may affect original behavior as well, but they are not of such overwhelming importance when the training of originality is considered. Our subsequent discussion therefore will be concerned solely with originality as here defined. One further distinction in the experimental study of original behavior needs to be made. Such behavior is always relative, either to a given individual's past behavior or to the norms of a population of which he is a member (Barron, 1955). The two are no doubt correlated. But it is the latter which is related to creative works; and since all of the pertinent experimental research has been of this kind, we shall be concerned only with the study of originality relative to some group norm.

---

[1] This work was supported by contract Nonr 233(50) between the Office of Naval Research and the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[2] These are only selected references and are not intended as an exhaustive survey of the relevant literature. A more complete list of references on originality may be found in Taylor (1956, 1958).

Guilford (1959) also distinguishes between originality and creativity. However, the basis for the distinction stems from his factor analytic approach. Originality, defined essentially in the same fashion as in the present paper, is one of several behavioral traits contributing to creativity. The latter is a more general behavioral trait including as components several forms of flexibility, fluency, and motivational and temperamental traits in addition to originality. Although this is a reasonable empirical classification of traits based upon factor analysis, it does not necessarily contradict the present one. We do not, however, accept Guilford's additional assertion that an S-R approach cannot deal adequately with creativity.

EARLIER ACCOUNTS OF ORIGINALITY TRAINING

Although psychologists have done relatively little experimental research on the training of originality, nonpsychologists have been concerned with the problem for some time. What is more, their suggestions, although not supported by evidence meeting the usual scientific standards, are frequently in accord with commonly accepted behavioral principles.

For example, Mearns (1958) in a book originally published in 1929 repeatedly emphasizes that in order to facilitate the originality of school children in the arts the teacher must reinforce, manifestly approve, his original efforts. He illustrates many techniques for inducing the child to make public his efforts, to gain the confidence of children so that communication will occur with their "secret unexpressed selves." Furthermore, the teacher is advised to approve of only the genuinely original effort. The teacher is further cautioned to wait patiently for the appearance of original behavior which is fostered by a "permissive atmosphere," the absence of "drill" and excessive discipline. According to Mearns, original behavior appears eventually because all normal children have an urge, energy, or impulse to be creative.

Aside from the romantic approach to a problem of behavior, we would agree that the way to foster originality is to reinforce such behavior when it occurs. A basic difficulty is that it may not occur at all or at such infrequent intervals that the reinforcements cannot shape up such behavior. Thus the fundamental problem in the training of originality is to devise methods for increasing its occurrence in diverse situations, is to get it to occur in the first place, thereby permitting the operation of reinforcement. If the operant level of originality is too low, conditioning will not be effective, training cannot occur. We are assuming, then, that originality can be learned and that the same principles of conditioning hold as in other forms of operant behavior. Once it occurs it can be reinforced, approved, recognized. It will then show an increased probability of occurrence. An apparent difficulty, however, is that originality is manifestly different from other behaviors. How can the reinforcement of one bit of uncommon behavior increase the frequency of other uncommon behaviors which by definition are different. Despite the apparent paradox, this sort of nonspecific transfer does occur, as we shall see.

Royce (1898), as far as we know, was the first to experimentally attack the problem of devising techniques for facilitating the occurrence of originality. He employed two different procedures. First, *S*s were asked to draw a series of nonsense figures unlike anything they had ever seen. A second procedure was to present a series of drawings to *S* with instructions to draw figures as different as possible from each of the models. Examination of the qualitative changes that occurred under these conditions led Royce to conclude that these procedures may facilitate originality and that they correspond to conditions in society that are conducive to originality.

Slosson and Downey (1922) suggest another method for training and testing originality in writing fiction. They suggest writing a plot, characters, or an entire story, based upon the more unusual messages found in the personal columns of newspapers. Headlines are also proposed as training materials. Variations are possible in the use of the materials by placing restrictions on the kind of plot or characters to be invented. Thus, given an unusual message, the writer is forced to practice literary invention. The assumption is that this form of "mental gymnastics" will increase the originality of his writing.

A simple test of originality suggested by Slosson and Downey (1922) is to determine how many different plots can be written to the same message, and for how many messages can plots be written in a given time interval. A more formalized test is also presented, but no empirical data are given from its use. Likewise, no data are presented on the effectiveness of the training method which appears to have been employed as a teaching technique in creative writing courses at the college level.

Osborn (1957) also believes that originality can be trained and that practice in producing original "ideas" will further develop originality. He gives many exercises designed to provide such training. They typically consist of various kinds of problem situations for which as many different solutions as possible are to be offered. The quantity of different "ideas" is stressed, since it is believed that the likelihood of original "ideas" increases with increases in quantity.

Osborn (1957) has also initiated a procedure of "group ideation" known as "brainstorming" which is designed to evoke original "ideas," and which has received considerable attention from the public press. It is a modified group free association procedure with emphasis upon producing large numbers of "ideas" in the absence of criticism or judgments as to their value. Critical evaluation of "brainstorming" sessions occurs later in order to avoid inhibiting unusual "ideas" during the session.

Although there have been courses established to teach "brainstorming" in business and industrial firms as well as colleges, there is as yet little objective research demonstrating its efficacy.

Taylor, Berry, and Block (1957) found that group brainstorming did not produce a greater number of "ideas" or unique "ideas" as compared to individuals working alone and whose performance was scored as a nominal group. In fact, it was found that the nominal groups were significantly superior to the brainstorming groups under the conditions of their experiment. Corroborative evidence was found in another study by Taylor and Block (1957). These experiments, however, were not designed to study the effects of originality training in groups or by individuals. They provide no information on the problem of training originality, whether giving many different solutions to problems results in an increased probability of giving original solutions to subsequent different problems.

Meadows and Parnes (1959) have recently reported that a course based upon the principles described in Osborn's book (1957) produced a significant increment in originality in its students as compared to a control class taking a different course. But as in many classroom experiments, there is the difficulty of assessing what the relevant variables are, whether it was the training in problem solving per se, or the changes in motivation resulting from participating in a class and studying a procedure which explicitly claims to facilitate problem solving. There are additional difficulties inherent in the study since no attempt was made to control or assess the extent of differential rehearsal of the test materials in experimental and control classes. The experimental class practiced on materials very similar to the originality tests given before and after the courses were taken. Finally, the design employed permits a biased sampling of *S*s, since the "creative problem solving" course was an elective one.

The early psychological literature as well as the suggestions of the nonpsychologists reviewed tend to agree upon a small number of different procedures for increasing originality. One training procedure is to present an uncommon stimulus situation, a situation for which common or conventional responses may not be readily available. Relatively uncommon responses may be evoked as a consequence (Slosson & Downey, 1922; Osborn, 1957). A second procedure is the evocation of different responses to the same stimulus situation (Royce, 1898; Osborn, 1957). Under such conditions the successive responses may become more uncommon. A third training procedure is the evocation of the uncommon responses as textual responses. Several recent experiments have employed this procedure (Judson, Cofer, & Gelfand, 1956; Maltzman, Brooks, Bogartz, & Summers, 1958; Maltzman, Simon, Raskin, & Licht, 1960) and will be discussed shortly. Instructions, as distinguished from training, may also be used to increase originality, and appear to be effective under certain conditions (Christensen, Guilford, & Wilson, 1957; Maltzman, Bogartz, & Breger, 1958; Royce, 1898).

"PRODUCTIVE" THINKING AND ORIGINALITY

Before turning to the current experimental research on originality training, however, note should be taken of several recent studies that have employed Maier's (1931) two-string or analogous problems as a test situation. The reason for commenting upon these studies is the assumption that there is no fundamental difference in the behavioral principles determining originality and problem solving behavior generally. Both involve the evocation of relatively uncommon responses, otherwise the situation would not be called a problem or the behavior original. This is especially apparent in connection with Maier's two-string problem. Maier (1931) considers the pendulum solution to be a consequence of productive thinking whereas other solutions represent reproductive thinking. As indicated elsewhere (Maltzman, 1955) these problem solving behaviors do not involve fundamentally different laws as demanded by Maier, but only a difference in the extent to which initial response hierarchies are modified. Furthermore, as far as can be determined this also is the basis that Maier used for distinguishing reproductive from productive solutions. The former solutions were given by his *S*s upon the initial presentation of the problem. Only after these solutions were given and the *S*s were told to find still another solution, which happened to be the pendulum solution, did Maier call it productive thinking. The empirical basis for the distinction between reproductive and productive thinking, then, was the probability of occurrence of the solution under his experimental conditions. The pendulum solution was relatively uncommon and is original in this sense.

Several studies have attempted to determine, with some success, whether verbal responses necessary for the solution of a subsequent problem are present in the initial response hierarchy of *S*s (Saugstad, 1955; Saugstad & Raaheim, 1957; Staats, 1957). The determination is made by obtaining word associations to objects later found in the problem situation. Responses necessary for the solution of the problem are presumably uncommon, have a low probability of occurrence in the initial response hierarchy. An association procedure has also been used to determine the uncommonness of responses evoked in the initial response hierarchy and their relation to subsequent success in the two-string problem (Maltzman, Brooks, Bogartz, & Summers, 1958).

A series of experiments by Judson, Cofer, and Gelfand (1956) is a limiting case of the use of textual responses in facilitating "productive" problem solving or originality. They employed the two-string problem as the test and rote learning as the training situation. Instead of attempting to increase the occurrence of uncommon responses during training and determining their transfer effects in the test situation, they induced directly the response chain necessary for a pendulum solution to the problem. A group form of the two-string problem was employed, and the *S*s were required to write as many solutions as possible to the problem of grasping both strings simultaneously. The design of the experiment was to prompt a pendulum solution to the problem by presenting an experimental group with different lists of words, one of which contained the relevant response chain "rope," "swing," "pendulum." Additional control groups were employed receiving word lists including none of these words, one or all three but each in a different list. In this study as well as in two replications, the experimental group was superior to the control groups, although the differences were not always statistically significant. A further qualification is that these results hold only for men. No reliable differences were obtained among women in the various conditions. An analogous experiment was conducted employing a group form of Maier's hat rack problem, and comparable results were obtained. These studies suggest that repeated evocation of the critical words during rote learning increased the probability of their occurrence in the problem situation which in turn lead to an increased frequency of the desired problem solution.

Another related study has been reported by Flavell, Cooper, and Loiselle (1958). Their study was designed to demonstrate that "functional fixedness" could be reduced by evoking uncommon responses to an object during a pretest period. In one respect it is more pertinent to the problem of producing relatively nonspecific transfer than the studies by Judson, Cofer, and Gelfand (1956), because the responses evoked during the pretest period were not the appropriate ones in the subsequent test problem. However, for our purposes the experiment suffers the limitation that it does not show an actual facilitation of problem solving performance but only a reduction in "functional fixedness."

A series of studies by Maltzman, Brooks, Bogartz, and Summers (1958) combines features of both of these experiments (Judson et al., 1956; Flavell et al., 1958) by evoking uncommon textual responses during a pretest period which are not involved in the problem solution in any obvious manner. Evidence of facilitation was obtained under these conditions. A screwdriver and a block of balsa wood were the objects available for use in the problem situation. Only the screwdriver was sufficiently heavy to serve as a weight permitting a string to swing in pendulum fashion. In the first experiment a control group was given the test problem without any pretraining. Lists of unusual uses for a screwdriver, balsa wood, and string were read by the experimental group before they performed in the problem situation. None of the uses referred to operations involved in the experiment. It was nevertheless found that the experimental group solved the problem significantly faster than the control group, and these results held for both sex groups. Additional experiments with modifications in the lists of unusual uses also obtained reliable differences between experimental and control *S*s, but typically in the case of women and not the men. Many variables in this experimental situation are in need of further analysis. However, the inherent lack of reliability of the two-string problem as a test and the difficulties involved in obtaining experimental control over the relevant variables in the situation suggest that more fruitful studies of originality and problem solving generally should employ another type of experimental situation. Also, with respect to the study of originality, test problems of this sort usually involve uncommonness of response only by inference or by definition. A more direct measure of originality would be highly desirable.

CURRENT STUDIES OF ORIGINALITY TRAINING

Because they seemed to meet the desired requirements, tests of originality of the kind developed by Guilford (1950) and his associates (Wilson, Guilford, & Christensen, 1953) as well as earlier writers (Hargreaves, 1927; Slosson & Downey, 1922) were chosen for a more detailed experimental analysis of variables involved in training originality by Maltzman and his associates (1958, 1960).

In the first experiment in this series (Maltzman, Bogartz, & Breger, 1958) the *S*s were run individually. All of the subsequent studies were conducted in group form. A modified free association procedure was employed in which the stimulus words were selected on the basis of their restricted response hierarchies. Since a relatively small number of different responses are evoked by the stimuli, a high degree of communality occurs among their responses. A control group received an initial list of 25 words to which they gave free associations and a different final list of 25 words to which they gave free associations. This was followed by the Unusual Uses Test of originality (Guilford, 1950). The latter test consists of a series of names of six common objects for which *S* must give different uses other than their common everyday use.

An experimental group received the same treatment except that the initial list was presented five additional times with instructions to give a different response on each repetition. A second experimental group was treated in the same fashion except that *E* said "good" after every approximately fifth uncommon response during training. Each of these three groups was subdivided prior to the introduction of the free association test list. Half the *S*s in each condition were instructed to try to be as original as possible, and half were told nothing. Each of these subgroups were further subdivided prior to administration of the Unusual Uses Test and instructed to be as original as possible. These latter instructions may have been redundant, if not confusing, since the written instructions to the test already implied as much.

Originality measures were obtained for performance on the free association list by determining the frequency with which the responses to each stimulus word occurred. The mean frequency of his associative responses was taken as each *S*'s originality score. Thus the lower the score the more uncommon are *S*'s responses on the average. The total number of unique responses given to the items was taken as the measure of originality on the Unusual Uses Test, where a unique response was defined as a use occurring only once in the sample. Results of the experiment showed that both training and instructions produced a significant increase in originality on the free association test list. What is more, the experimental subgroups receiving training in addition to instructions were significantly more original than the control group receiving instructions alone, indicating that the training may produce its effects independently of instructions to be original. Results obtained from the Unusual Uses Test were ambiguous, yielding a significant triple order interaction where the experimental subgroup without originality instructions was significantly more original than its control group. However, results on the Unusual Uses Test from subsequent experiments have consistently shown reliable training effects.

Despite the above evidence indicating that the training procedure facilitated originality, additional information is needed to determine precisely what characteristics of the procedure are responsible for the effect. Differences between the experimental and control groups may have been due to the difference in number of stimulus words presented, number of responses evoked, number of different responses evoked, etc. The fact that different responses were evoked by the same stimuli may be neither necessary nor sufficient for an effective training method. Additional research was therefore conducted in an attempt to isolate the relevant variables.

In one study (Maltzman et al., 1960) the control condition and the experimental condition without verbal reinforcement or instructions to be original were again employed. The latter group received instructions to give a different response to each presentation of the same stimulus word. For convenience they will be called the standard control and experimental conditions, respectively. Two additional experimental groups received lists of 125 different stimulus words, the same number as the repetitions of the initial list for the standard experimental group. One group received high frequency count stimulus words while the other group was presented with relatively low frequency count words (Thorndike & Lorge, 1944). This procedure equated the experimental groups for the total number of different responses. They differed in that the responses all occurred to different stimulus words in the new experimental conditions whereas different responses occurred to the same stimulus words in the standard experimental condition. Another control group was added which received the same number of repetitions of the initial list as did the standard experimental group, but these *S*s were instructed to try to give the same response at each repetition of a given stimulus. This condition would aid in determining whether the number of responses per se, or the number of different responses, is the relevant variable influencing test performance.

Results from the free association test of originality showed that the three experimental groups did not differ significantly from each other and the two control groups did not differ significantly. But each of the experimental groups was significantly more original than each of the control groups. On the Unusual Uses Test the standard experimental group did significantly better than all the other groups. The two experimental groups receiving different stimulus words during training did not differ from the standard control group, while the new control condition was significantly less original than each of the other conditions.

The results from this experiment indicate that under some conditions the evocation of a relatively large number of different responses to different stimuli may increase originality to the same extent as the method of repeated evocation of different responses to the same stimuli. However, the two methods apparently differ in the extent to which they produce transfer effects, as indicated by the significant difference between the two methods on the Unusual Uses Test. These results show that the evocation of different responses to the same stimuli is an important aspect of the standard experimental procedure, although arousal of many different responses by different stimuli is also effective, within limits.

The significant loss of originality by the new control group was not entirely unexpected, and shows that a training technique can significantly decrease as well as increase originality, a result which is of theoretical interest as well as suggestive for classroom practices. Exactly why the effect should appear on the Unusual Uses Test and not on the preceding free association test is not clear, and is a problem that requires further experimental investigation.

Another experiment (Maltzman et al., 1960) was designed to further explore possible relevant variables and alternative procedures for facilitating originality.

The standard control and experimental conditions were again employed. Another experimental condition was given repeated presentation of items from the Unusual Uses Test in order to determine whether the effectiveness of the training procedure is limited to only certain kinds of materials. This was accomplished by presenting the list of six items a total of six times with instructions to give a different use prior to each repetition. A different kind of training procedure was employed with a third experimental group. They were presented with a booklet consisting of 125 different pairs of words. The *S*s were instructed to indicate the member of each pair which goes more readily with the stimulus word orally presented to them. The same stimulus words and number of repetitions as employed with the standard experimental conditions were administered. Each pair of words were unique responses to the stimulus words selected from the previously established norms. A fourth experimental group was given the word pairs and asked to indicate which member of each pair they thought to be more familiar. The stimulus words were not presented. Results of the free association and Unusual Uses Tests showed that the standard experimental condition was significantly more original than the other four groups on both tests, while the other groups did not differ significantly from each other.

In one of the new conditions uncommon responses were evoked as textual responses following the stimulus words. In the other condition only the textual responses occurred. These conditions were introduced in order to determine whether the uncommon responses must be evoked by the stimulus words in order for the training to be effective. The fact that these groups were significantly less original than the standard experimental group suggests that, at least with the training and test materials employed, uncommon responses must be evoked as intraverbal responses by the stimulus words if there is to be an increase in originality in the test situations.

Since the experimental condition which received original training with items from the Unusual Uses Test failed to show a significant training effect, another experiment was conducted to determine whether the failure was due to an inadequate amount of training. Four conditions were employed in this study receiving different amounts of training. None of the groups showed a significant increase in originality on the free association test when compared with the standard control condition.

Results from the latter experiment and one condition from the previous experiment showed that training with unusual uses items in the same fashion as word association materials does not produce a comparable increase in originality. Thus the method of repeated evocation of different responses to the same stimuli does not work with all materials.

A basic assumption of our approach is that, if original behavior can be induced to occur repeatedly through some training procedure, it will increase in its frequency of occurrence in new situations. If this condition is not met, transfer of originality to new situations cannot occur. Inspection of the unusual uses training data indicated that an appreciable increase in originality did not occur during training. In the case of training with free association materials a marked increase does take place.

Another experiment was conducted in which different groups received 0, 1, 5, or 10 repetitions of the initial word association list before testing on another word association list and the unusual uses. A significant trend was evidenced on both originality tests, with a single repetition of the training list producing a significant increase in originality (Maltzman et al., 1960).

On the basis of the current dearth of relevant information, we can only speculate as to why training with unusual uses failed to induce an increase in the uncommonness of the responses. One hypothesis that may be offered is based upon the apparent relatively limited number of different uses available for each item. Given the instructions to write a different use to each repetition of an item and the inability of *S*s to do so with a different operant, they respond with essentially synonymous uses repeatedly, a condition, as we have already seen, not likely to facilitate originality.

If, as we have assumed, originality can be learned according to the principles of operant conditioning, then the effects of originality training should be of the same kind as those ordinarily obtained in experimental studies of learning. One of the gross characteristics of learning is that the behavioral changes produced by reinforcement persist for some time, are not relatively transistory. Increased originality induced by the training method previously described should therefore also persist. An experiment designed to investigate this property of originality training has been conducted, and positive results have been obtained (Maltzman et al., 1960). Using the standard control and experimental conditions of the previous experiments, a delay of approximately one hour was interposed between the training and test situations for a control and an experimental group, and a delay of two days for another control and experimental group. A significant training effect was obtained on the word association and the Unusual Uses Test of originality.

These results show that the effects of originality training persist to a significant degree for at least two days, under the given experimental conditions. We may conclude from these results and that of the previous experiment that the standard experimental procedure for training originality produces some kind of learning.

RELATION OF ORIGINALITY TRAINING TO LEARNING TO LEARN

Originality training studies of the kind reviewed here and learning to learn studies (Duncan, 1958; Harlow, 1949; Morrisett & Hovland, 1959) have the common characteristic of producing relatively nonspecific transfer effects. Training on given stimuli transfers to other stimuli in the absence of any obvious opportunity for the operation of primary stimulus generalization. Despite the apparent similarity between the results of training in learning to learn and those discussed here, differences in the experimental operations and the kind of behavior induced in these procedures suggest that different principles are involved in these situations.

The learning to learn procedure involves the learning of a great many different discrimination problems. Different stimuli are presented repeatedly until *S* has learned a differential response to a criterion or has received a given number of trials. The responses are not varied and uncommon responses are not reinforced, nor can they be, because a specified response is correct. Under these conditions *S*s show a progressive decrease in the number of trials required to learn new discrimination problems. In originality training, on the other hand, variability of response is induced. Repeated presentation of the same stimuli and reinforcement of a specific predesignated response likewise does not occur in originality training. In learning to learn, the stimuli vary from problem to problem, but the responses do not. In originality training and test situations the stimuli and the responses vary. Thus the two procedures differ in terms of the operations employed to induce nonspecific transfer. Furthermore, the interpretations offered by Duncan (1958) in his comprehensive study of learning sets at the human level do not appear applicable to originality training. The acquisition of observing responses, habits of looking at details of stimuli, etc. are important in discrimination problems but do not appear to be particularly relevant in originality training situations of the kind described here. Likewise, the analysis of learning to learn by Morrisett and Hovland (1959) in terms of the discrimination between stimulus trace patterns does not appear applicable to originality training.

AN INTERPRETATION OF ORIGINALITY TRAINING

Two basic problems, not unrelated, posed by the experiments on originality training by Maltzman and his associates (1958, 1960) demand an interpretation. First is the nature or source of the reinforcement for originality. In all of the experiments, successive training trials in the standard experimental conditions are accompanied by increases in the uncommonness of the responses. The originality induced in this manner transfers to different stimulus materials, and these effects tend to persist. Originality training thus produces behavioral changes that are characteristic of learning, yet they occur in the absence of differential reinforcement administered by *E*. Nevertheless, we believe that uncommon responses received differential reinforcement. Reinforcement was obtained, because the occurrence of any previously established intraverbal association is self-reinforcing. As Thorndike puts it: "The mere occurrence of a connection strengthens it" (1949, p. 23). Thus, where there is an initial probability greater than zero that a given verbal stimulus will evoke a verbal response in an operant situation such as word association, the occurrence of that response will increase the probability of its occurrence on a subsequent presentation of that stimulus. Different initial conditions may only increase or decrease the amount of reinforcement affecting the stimulus-response connection. In the type of experimental situation with which we are concerned here, the uncommonness of the response evoked by the stimulus is one of these conditions. In other words, the amount of reinforcement is inversely related to the initial probability of a verbal stimulus evoking a verbal response. Originality is more reinforcing than commonplace responding.

An impression gained from observing *S*s in the experimental situation is that repeated evocation of different responses to the same stimuli becomes quite frustrating: *S*s are disturbed by what quickly becomes a surprisingly difficult task. This disturbed behavior indicates that the procedure may not be trivial and does approximate a nonlaboratory situation involving originality or inventiveness, with its frequent concomitant frustration. This impression of frustration accompanying the task of producing repeated uncommon responses would lend support to the notion that when such responses do occur they are self-reinforcing to a considerable degree. These speculations, however, are not essential to the investigation of the initial hypothesis of the self-reinforcing character of uncommon responses. However, these as well as other considerations suggest that the distribution and scheduling of the evocation of uncommon responses may be of great importance in determining the extent to which a transfer effect will be obtained.

It should be noted, however, that the role of self-reinforcement is probably less significant in producing an increase in originality during training, the successive repetitions of the same word list, than in the test situation. Since instructions to the standard experimental group prohibits the reoccurrence of the same response to a given stimulus word, successive responses to that word will be progressively lower in the related response hierarchy. This is sufficient to account for the increase in the uncommonness of responses with successive repetitions of the same stimulus words. However, we believe that a second variable contributes to the increase in originality during the training session, generalization of the reinforcement effects from the evoked response to still more uncommon responses in the same hierarchy. However, the role of self-reinforcement becomes significant when the transfer of originality to the test situations is considered. In this case the carrying out by *S* of previously administered instructions to give different responses will not account for the greater originality of the standard experimental condition in the test situations.

This leads to the second problem presented by the experimental data: the fact that the evocation of uncommon responses on one list of stimulus words facilitates the occurrence of original responses on a different list of words, and even more perplexing, on the rather different Unusual Uses Test. The problem is to account for the fact that reinforcement of one instance of original behavior increases the response tendency for other instances of such behavior under different stimulus conditions. A relatively nonspecific kind of transfer of training is induced. Without it, the facilitation of originality would not be possible.

We do not have any simple rigorous explanation of this phenomenon at the present time, but offer the following tentative suggestions. The intraverbal associations possessed by a normal educated adult are enormously extensive and complex. In all likelihood, almost every verbal response is associated to some extent with every other. Furthermore, the intraverbal associations among common verbal responses are stronger than between common and uncommon responses. Likewise, the intraverbal associations among uncommon responses are relatively stronger than between uncommon and common. The evocation and reinforcement of uncommon responses will, through complex kinds of mediated generalization, therefore increase differentially the probability of occurrence of other uncommon responses even though they do not ordinarily occur to the same stimuli as the original reinforced responses. In a similar fashion, the evocation of common responses would differentially increase the probability of occurrence of other common responses even though they may be members of different response hierarchies. An additional characteristic of the standard experimental procedure used in originality training is that it provides for the inhibition of common responses as well as the facilitation of uncommon responses. Through mediated generalization the effects of inhibition will produce a decrement in the excitatory potential of other common responses. It is this characteristic which may be responsible for the more general transfer effects obtained with the standard experimental procedure than with the procedures that induce uncommon textual responses or employ uncommon stimulus situations for evoking, in turn, uncommon responses. None of these other procedures would involve the inhibition of common responses.

It should be clear that the foregoing hypotheses are speculative and their connection with the data presented is tenuous. Considerable careful experimentation tracing the intraverbal associations present in the experimental situation is needed as well as the development of laws of verbal compounding before a completely adequate explanation of the basis for originality training can be given. Whether or not such explanations can be given, the method of training originality described in this paper seems worthy of further study in its own right, and potentially may have considerable practical application. This again is a problem requiring sustained experimental research employing different kinds of test materials.

In conclusion, it should be emphasized that little is to be gained by disputing the usage of the term "originality" as it has been employed here, or whether the current experiments described are investigating the essence of originality. The important problem is the experimental one of determining the variables influencing the occurrence of uncommon responses in relatively simple situations and the functional relationships that obtain in these situations. Finally, the extent to which these effects hold for more complex behavioral situations must be determined. Such work is now in progress in the UCLA laboratory. But to designate only highly complex behavior of a given sort as original to the exclusion of simpler better controlled behaviors seems gratuitous, as well as an obstacle to the progress of research in this vital area.

SUMMARY

The basic problem in the training of originality is to devise a means of increasing the frequency of uncommon behavior. Once it takes place it may receive reinforcement and increase the probability that other original behavior will occur. Earlier attempts to devise training methods for originality were briefly mentioned, as well as related studies of problem solving.

A series of experiments by Maltzman and his associates was reviewed, and a procedure which consistently facilitated originality was described. This procedure involves the repeated presentation of a list of stimulus words in a modified free association situation accompanied by instructions to give a different response to each stimulus. Under these conditions the responses become more uncommon. When presented with new stimulus materials, *S*s receiving such training are reliably more original than *S*s receiving no training. Tentative suggestions as to the behavioral bases for the training effect were given.

REFERENCES

BARNETT, H. G. *Innovation: The basis of cultural change.* New York: McGraw-Hill, 1953.

BARRON, F. The disposition towards originality. *J. abnorm. soc. Psychol.,* 1955, **51,** 478–485.

CHRISTENSEN, P. R., GUILFORD, J. P., & WILSON, R. C. Relations of creative responses to working time and instructions. *J. exp. Psychol.,* 1957, **53,** 82–88.

DENNIS, W. The age decrement in outstanding scientific contributions: Fact or artifact? *Amer. Psychologist,* 1958, **13,** 457–460.

DREVDAHL, J. E. Factors of importance for creativity. *J. clin. Psychol.,* 1956, **12,** 21–26.

DUNCAN, C. P. Transfer after training with single versus multiple tasks. *J. exp. Psychol.,* 1958, **55,** 63–72.

FLAVELL, J. H., COOPER, A., & LOISELLE, R. H. Effect of the number of preutilization functions on functional fixedness in problem solving. *Psychol. Rep.,* 1958, **4,** 343–350.

GHISELIN, B. (Ed.) *The creative process.* New York: New American Library, 1955.

GUILFORD, J. P. Creativity. *Amer. Psychologist,* 1950, **5,** 444–454.

GUILFORD, J. P. Traits of creativity. In H. H. Anderson (Ed.), *Creativity and its cultivation.* New York: Harpers, 1959.

HARGREAVES, H. L. The faculty of imagination. *Brit. J. Psychol., Monogr.,* 1927, No. 10.

HARLOW, H. F. The formation of learning sets. *Psychol. Rev.,* 1949, **56**, 51–65.

JUDSON, A. J., COFER, C. N., & GELFAND, S. Reasoning as an associative process: II. "Direction" in problem solving as a function of prior reinforcement of relevant responses. *Psychol. Rep.,* 1956, **2**, 501–507.

LEHMAN, H. C. *Age and achievement.* Princeton: Princeton Univer. Press, 1953.

MAIER, N. R. F. Reasoning in humans: II. The solution of a problem and its appearance in consciousness. *J. comp. Psychol.,* 1931, **12**, 181–194.

MALTZMAN, I. Thinking: From a behavioristic point of view. *Psychol. Rev.,* 1955, **62**, 275–286.

MALTZMAN, I., BOGARTZ, W., & BREGER, L. A procedure for increasing word association originality and its transfer effects. *J. exp. Psychol.,* 1958, **56**, 392–398.

MALTZMAN, I., BROOKS, L. O., BOGARTZ, W., & SUMMERS, S. S. The facilitation of problem solving by prior exposure to uncommon responses. *J. exp. Psychol.,* 1958, **56**, 399–406.

MALTZMAN, I., SIMON, S., RASKIN, D., & LICHT, L. Experimental studies in the training of originality. *Psychol. Monogr.,* 1960, **74**(6, Whole No. 493).

MEADOW, A., & PARNES, S. J. Evaluation of training in creative problem solving. *J. appl. Psychol.,* 1959, **43**, 189–194.

MEARNS, H. *Creative power: The Education of youth in the creative arts.* New York: Dover, 1958.

MORRISETT, L. JR., & HOVLAND, C. I. A comparison of three varieties of training in human problem solving. *J. exp. Psychol.,* 1959, **58**, 52–55.

OSBORN, A. F. *Applied imagination.* New York: Scribner's, 1957.

ROYCE, J. The psychology of invention. *Psychol. Rev.,* 1898, **5**, 113–144.

SAUGSTAD, P. Problem-solving as dependent on availability of functions. *Brit. J. Psychol.,* 1955, **46**, 191–198.

SAUGSTAD, P., & RAAHEIM, K. Problem-solving and availability of functions. *Acta Psychol.,* 1957, 13, 263–278.

SLOSSON, E. E., & DOWNEY, J. E. *Plots and personalities.* New York: Century, 1922.

SPRINGBETT, B. M., DARK, J. B., & CLARKE, J. An approach to the measurement of creative thinking. *Canad. J. Psychol.,* 1957, **11**, 9–20.

STAATS, A. W., Verbal and instrumental response hierarchies and their relationship to problem-solving. *Amer. J. Psychol.,* 1957, **70**, 442–446.

STEIN, M. I. Creativity and culture. *J. Psychol.,* 1953, **36**, 311–322.

TAYLOR, C. W. (Ed.) *The 1955 University of Utah research conference on the identification of creative scientific talent.* Salt Lake City: Univer. of Utah Press, 1956.

TAYLOR, C. W. (Ed.) *The second (1957) University of Utah research conference on the identification of creative scientific talent.* Salt Lake City: Univer. Utah Press, 1958.

TAYLOR, D. W., BERRY, P. C., & BLOCK, C. H. *Does group participation when using brainstorming facilitate or inhibit creative thinking? ONR 150-166, Tech. Rep. 1* 1957, Dept. Industr. Admin. and Dept. Psychol., Yale Univer.

TAYLOR, D. W., & BLOCK, C. H. *Should group or individual work come first on problems requiring creative thinking when equal time is devoted to each? NR 150-166, Tech. Rep. 1* 1957, Dept. of Indust. Admin. and Dept. Psychol., Yale Univer.

THORNDIKE, E. L. *Selected writings from a connectionist's psychology.* New York: Appleton-Century-Crofts, 1949.

THORNDIKE, E. L., & LORGE, I. *The teacher's word book of 30,000 words.* New York: Columbia Univer. Press, 1944.

VAN ZELST, R. H., & KERR, W. A. Some correlations of technical and scientific productivity. *J. abnorm. soc. Psychol.,* 1951, **46**, 470–475.

WILSON, R. C., GUILFORD, J. P., & CHRISTENSEN, P. R. The measurement of individual differences in originality. *Psychol. Bull.,* 1953, **50**, 362–370.

(Received September 4, 1959)

Psychological Review
1960, Vol. 67, No. 4, 243–252

PREPARATORY SET VARIABLES RELATED TO CLASSICAL CONDITIONING OF AUTONOMIC RESPONSES[1]

WILLIAM W. GRINGS

University of Southern California

Experiments on conditioning with human subjects (*S*s) are complicated by the fact that response change or "learning" can take place at different levels, sometimes facilitating and other times interferring with the experimenter's (*E*'s) goals. Razran (1955) emphasized this point when he distinguished among five relations of conditioning and perception. Of his categories two are particularly relevant here, representing a distinction between "conditioning without perception" and "conditioning with perception," the difference depending largely upon whether the relation between the unconditioned reaction and the to-be-conditioned reaction is perceived by *S*.

As a specific example consider studies conditioning the GSR or change in heart rate with human *S*s. Over twenty years ago Cook and Harris (1937) demonstrated one trial "verbal conditioning" of the GSR in which the strength of the response was not dependent upon the number of reinforcement trials. They concluded, not only that the GSR is verbally conditioned, but also that such "verbal conditioning is impossible to prevent in conditioning by the usual procedure." In spite of this early result, most GSR "conditioning" studies since that time appear to have been conducted either ignoring verbal-perceptual factors or assuming that they are controlled by instructions designed to mislead the *S*s concerning the nature of the experiment.

When the "perceptual" problem has been attacked directly the favored approach has been to manipulate verbal and attitudinal conditions through changes in instructions to *S*, informing him of stimulus relations or asking him to take a voluntary part in the process. Early studies of this type were summarized by Hilgard and Marquis (1940). Later studies have not emphasized autonomic responses, although Lacey and Smith (1954) and Lacey, Smith, and Green (1955) demonstrated heart rate conditioning differences between "aware" and "unaware" *S*s; and Grings and Kimmel (1959) noted marked instruction effects in GSR discrimination learning. References to the evanescent nature of GSR conditioned responses are common (Grant & Schiller, 1953, Grings & Shmelev, 1959), and the implication is usually made that the GSR is sensitive to transient changes in human *S*s.

The above is ample to suggest that the handling of verbal perceptual set factors in human autonomic conditioning remains a problem of major importance. Data available are limited to demonstrations of the existence of perceptual factors, with very few attempts to define operations for their detection, evaluation, and control. Extension of understanding of the conditioning process with human *S*s may result from exploration of new operations for inferring and manipulating perceptual sets in situations resembling classical conditioning.

---

[1] This paper is an outgrowth of a program of research supported by grants from the National Science Foundation.

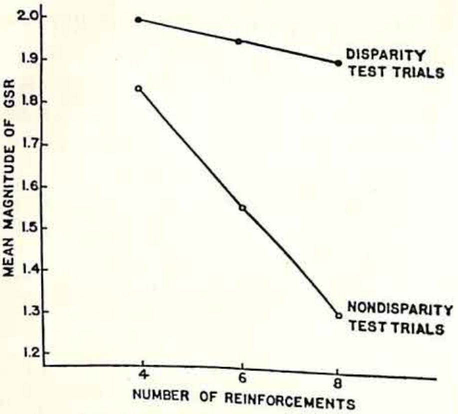


FIG. 1. Average responses on disparity trials and nondisparity trials for groups with different numbers of prior reinforcements. Tone and light stimuli. Response magnitude in square root of conductance change in micromhos.

THE PERCEPTUAL DISPARITY RESPONSE

The present research introduces an operation for inferring the existence of a perceptual set in conditioning studies, one that is not based on verbal instructions to the *S*. The concept is based on the premise that human *S*s, in situations with conditioned stimuli preceding noxious stimulation, use the cue as a preparatory signal—that is, the response to the conditioned stimulus (or what is, in effect, the conditioned response) is partly a preparatory set for receipt of the unconditioned noxious stimulus.

It is proposed that the strength of a perceptual set aspect of conditioned responses can be estimated from perceptual disparity situations, or from situations in which a cue stimulus is followed by an unconditioned stimulus different from that which had been associated with that cue previously. In other words, it is assumed that there will be a difference in magnitude of response between situations where receipt of stimulation is in accord with past experience and where such receipt of stimulation is not in accord with past experience. It is further assumed that a perceptual disparity variable, so defined, will be related to meaningful determining conditions in human behavior and lead to explanatory possibilities which go beyond more conventional definitions of conditioned responses.

To test the feasibility of such a disparity variable, several experiments have been conducted and will be described briefly below along with later studies amplifying upon the concept. In one experiment, three groups of six *S*s each were taught that when they moved a lever in one direction they would receive a tone stimulus and when they moved the same lever in the opposite direction they would receive a bright light. The groups were given different numbers of training trials (four, six, and eight) during which the above regularities applied. Then they were administered a test series containing both disparity and nondisparity trials. Disparity was achieved by following with a tone the direction of movement which previously had been followed by light, or by following with

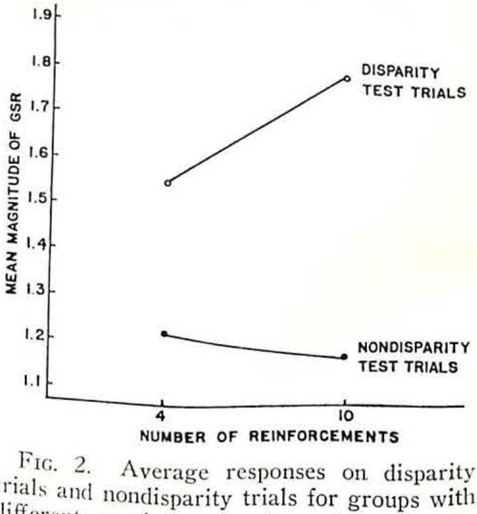


FIG. 2. Average responses on disparity trials and nondisparity trials for groups with different numbers of prior reinforcements. Tone and electric shock stimuli. Response magnitude in square root of conductance change in micromhos.

a light the direction of movement which had previously been followed by a tone.

The results from the test series are plotted in Figure 1 where it can be seen that the magnitude of response on disparity trials was larger than the response on the nondisparity trials for all groups and the difference between the curves increases with larger numbers of training trials. The difference between the disparity and nondisparity responses tested overall by a *t* test was significant at the 1% level; the effect of the number of reinforcements variable tested as an analysis of variance of the subgroup mean differences between disparity and nondisparity responses was not significant.

A second, similar study (Grings & Kimmel, 1958) was conducted using tone and shock stimuli rather than tone and light as UCSs. Sixty *S*s were observed, 30 being given 4 reinforcements before test and 30 being given 10 reinforcements before test. Again the test series contained both disparity and nondisparity trials. The results, shown in Figure 2, agree with those in the first study, i.e., for both groups the response (GSR) on disparity trials is larger than on nondisparity trials (significant at the 1% level) and the size of the difference between the disparity and nondisparity responses increases as a function of the number of reinforcements.

On the basis of these and related studies it was felt that the assumption of a disparity variable was reasonable, and attention was turned to the manipulation of the variable by introducing appropriate independent variation to which it might be related.

DETERMINERS OF THE PERCEPTUAL DISPARITY RESPONSE

On the basis of the already cited studies it was hypothesized that at least two major classes of determiners of the disparity response exist. One involves the strength and stability of the anticipatory response, which in turn appears to be a function of the number of reinforcements of the cue-consequent relations. If preparatory set responses increase and stabilize as most habits do, the more frequent the sequence cue-reinforcing stimulus the more strong and consistent should be the organism's preparation for the reinforcing stimulus, other variables held constant. Elaboration of this point, however, needs to be made, for the operation of competing processes, like response adaptation, is clearly evident in Figures 1 and 2. Further discussion of this matter is made later. In addition, numerous factors related to the conditioned stimulus used (such as its intensity, discriminability, etc.) probably determine the strength of the conditioned preparatory response.

The second major class of determiner of the magnitude of the disparity response would include the direction of the disparity—i.e., whether actual stimulation was stronger or weaker than that for which *S* was prepared—and the amount of disparity measured on some scale of stimulus distance from anticipated to actual stimulation.

Several studies have been conducted to explore the hypotheses outlined above, two examples of which are given below. One of these has been reported by Kimmel (1960). He was testing the relation between amount of classical conditioning of the GSR and the intensity of the CS. Tone CSs of intensities of 35, 55, 75, 95, and 115 db. re: .0002 dyne/cm^2 were used with an electric shock UCS. In attempting to control differences in response strength to cue stimuli during an extinction series he subdivided each acquisition CS intensity group into three extinction intensity subgroups (45, 75, and 105 db.). Since at the time of the first

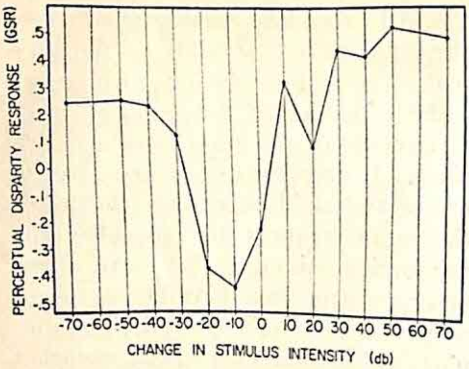


FIG. 3. Magnitude of perceptual disparity response as a function of amount and direction of stimulus change. (Reproduced from Kimmel, 1960.)

extinction trial *S* had just been given 20 trials during which the first stimulus he received on each trial was the acquisition intensity, it may be assumed that between trials he remained prepared for that intensity. When, therefore, he received an intensity different from the previously experienced one, he might respond to the disparity. Kimmel's data are presented in Figure 3 in terms of GSRs resulting from different amounts and directions of change. The effects of both direction and amount of change were reported as statistically significant.

Similar results were obtained in a study by Grings, Dossett, and Honnard (1959). Forty-eight college students were randomly separated into two experimental groups. One group was given conditioning training with 100% reinforcement, the other with 67% reinforcement. Two cue stimuli (colored light CSs) were followed by a 50-db. tone or a tone of different intensity. Four subgroups of 6 *S*s each were trained to different second stimuli of either 30, 70, 90, or 110 db. re: .0002 dyne/cm^2. A standard number of eight reinforcements of each cue-tone combination was followed by a disparity trial. Since each disparity trial was made up of the cue to the 50-db. stimulus followed by the actual "other" stimulus, there resulted four different degrees of discrepancy: -20 db., 20 db., 40 db., and 60 db.

The results of this experiment are presented in Figure 4 where mean response to the UCR component of the disparity trial (adjusted for reference level of the UCS involved) is given as a function of amount and direction of disparity. When an analysis of variance was made on the data shown, the difference between the 100% and 67% groups was found to be significant beyond the 5% level ($F = 6.80$, $df = 2$ and 6). The difference between the disparity subgroups, as based on the values portrayed by the combined groups curve in the figure, fell just short of the 5% level of significance ($F = 2.25$, $df = 4$ and 6).

On the basis of the above studies, it is tentatively concluded that the greater the amount of discrepancy between anticipated and actual stimulation the

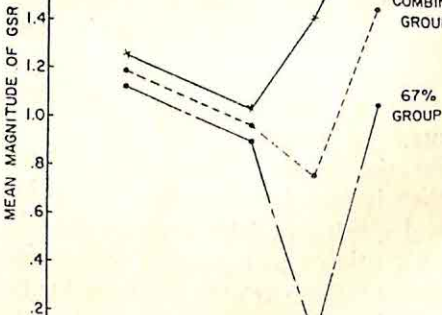


FIG. 4. Mean responses to the UCR component of the disparity trial, adjusted for reference level of the UCS involved, for groups differing in the percentage of prior trials reinforced.

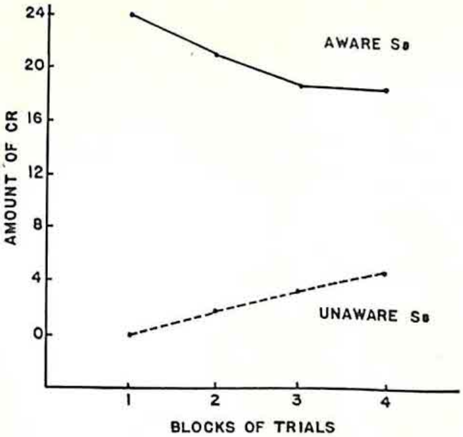


FIG. 5. Comparison of conditioned heart rate responses for groups differing in awareness of CS–UCS relations. (Adapted from Lacey, Smith, & Green, 1955.)

greater the GSR elicited by the actual stimulus. Discrepancies in an upward direction appear to yield larger disparity responses than discrepancies in a downward direction. In addition, these factors appear to be a function of the number of reinforcements and the percentage of reinforcements employed.

APPLICATIONS OF THE DISPARITY CONCEPT

Thus far the emphasis has been upon the definition of the disparity response and the exploration of some of its determining conditions. Next, attention is paid to application of the disparity situation in attempts to clarify some problems of human conditioning.

One of the problems that has concerned the writer is the nature of the GSR acquisition "learning" curve. In other words, what is the effect of the number of trials (or reinforcements) on the magnitude of a conditioned autonomic response? If the conditioned GSR, for example, is determined by interactions of different processes (i.e., nonperceptual and perceptual), it would be predicted that at times one process or the other might be dominant. One result could be that a "typical" learning curve would not be obtained: that is, GSR magnitudes as a function of number of reinforcements would not be an increase in CR magnitude with an increase in trials.

That different "learning" curves might be expected from different perceptual situations has been demonstrated by Lacey, Smith, and Green (1955). It will be noted from Figure 5 that they found "unaware" *S*s gave increasing heart rate CRs with more trials, whereas "aware" *S*s showed decreasing CRs with more trials. Similarly, Kimmel (1959), with GSR conditioning in this laboratory, found magnitude of anticipatory CRs to increase up to eight reinforcements, then to decrease. His data are given in Figure 6. The implication in both cases is not that less "learning" is occurring with more reinforcements but rather that the perceptual aspects of the situation are probably changing.

If it might be presumed that perceptual set factors are stabilized by repeated trials where a given CS–UCS relation obtains, it might be proposed that as long as no irregularities occur in the training situation *S* will cor-

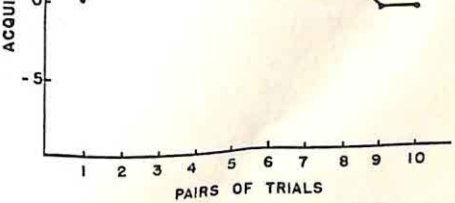


FIG. 6. Acquisition scores for GSR anticipatory CRs in successive blocks of two trials. (Reproduced from Kimmel, 1959.)

rectly "set" for the noxious UCS and therefore will show a curve of adaptation with successive stimulation (just as he would to a series of presentations of the UCS alone). The result will be that anticipatory CRs and UCRs may both show reduced amplitude with increased trials.

It might be argued further that it is the perceptual tendencies or sets which are becoming stronger with increased experience or reinforcements. If the disparity response is taken as evidence of perceptual set strength, it can be predicted from data already at hand (Figures 1 and 2) that the disparity variable (the *difference* between the response magnitudes on disparity and nondisparity trials) does increase with larger numbers of reinforcements. It was also found (Figure 4) that the disparity variable was greater for 100% than for 67% reinforcement conditions.

The proposal here, then, is that the disparity response may show a different course of development (with regard to the number-of-reinforcements variable) than will a simple anticipatory CR. Depending upon conditions it may provide a "better" estimate of strength of preparatory set than the anticipatory CR does. In order to examine this possibility, consider a current experiment in which an attempt was made to emphasize the perceptual set properties of the conditioning environment.

Taking a lead from Cornbecker, Welch, and Fisichelli (1949) the experiment was designed to test the notion that reinforcing a class of cues intrinsically suggesting stimulation will lead to increased perceptual response tendency (to expect stimulation) on a cue not reinforced but capable of suggesting stimulation. The hypothesis was tested by presenting a series of seven-word stimuli which were appropriately reinforced during the acquisition period, i.e., the words "soft music" were followed by a few bars of soft orchestral music, the words "cool breeze" were followed by a brief stimulation by an electric fan, "loud tone" followed by actual tone, "green light" by actual green light, etc. After either 14 or 49 such "reinforcements" a critical word or neutral word was introduced in a test series which also included a disparity trial (cue "green light" followed by the actual loud tone).

Since in all cases the cue was presented for 5 seconds before the reinforcing (UCS) stimulus was presented, it became possible to trace the course of anticipatory responses following any particular cue. Also, since long and short reinforcement series were employed, the effect of different numbers of reinforcements on both the regular and the disparity trials could be observed. The previous development would lead to the prediction that anticipatory responding (to the cue "loud tone") would become less with more reinforcements, as would also the response to the actual tone, due to adaptation. On the other hand, the disparity response to tone following another ("green light") cue would increase with greater numbers of reinforcements.

The results for this comparison for 20 *S*s receiving 14 reinforcements prior to the disparity trial and 20 *S*s receiving 49 reinforcements prior to the disparity trial are shown in Figures 7 and 8. The data follow the prediction closely, the anticipatory responses showing significant decrease from short to long reinforcement series ($t = 2.64$, $df = 38$) and the disparity responses showing a significant increase over the same period ($t = 2.23$, $df = 38$).[2]

---

[2] Essentially the same relations were obtained with changes in digital blood flow as the response variable.

Explanation of the above results has proceeded from familiar behavioral postulates concerning reinforcement and perceptual sets. The concept of reinforcement designated an empirical state of affairs with specified temporal regularities obtaining between classes of stimuli (CSs followed by UCSs). Similarly, the perceptual set concept was related to such general notions as orientative-investigative reactions of Pavlov (1927), anticipatory goal responses of Hull (1930), and the S–R form of Tolman's expectancy (MacCorquodale & Meehl, 1954). It was later pointed out to the writer that experimentation on disparity variables using different descriptive terms had been carried out in the laboratory of Anokhin in Russia. This led to further considerations, as follows.

Anokhin (1955) was interested in a general theory of adaptive behavior which would help to integrate the activities of physiologists and psychologists. One of his central notions is that of "reverse afferentation," a process of reverse signalization or feedback which follows each reflex act "like an echo" and which informs the organism whether the adaptive effect appropriate to the situation has been achieved or not. This afferent process provides the necessary conditions to the organism's stopping activity or passing on to other acts comprising the chain of adaptation. The notion of "reinforcement" can be applied to any fractional stage of adaptation which has obtained corresponding reverse afferentation. He talks about "sanctioning afferentation" which stops further attempts at organization of new reflex acts and strengthens the last combination of excitation in the centers of the brain which gave a successful adaptive effect at the periphery.

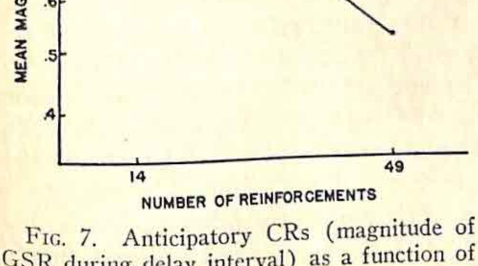


Fig. 7. Anticipatory CRs (magnitude of GSR during delay interval) as a function of the number of reinforced trials.

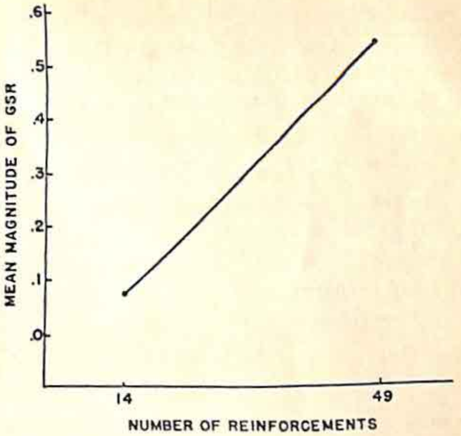


Fig. 8. Magnitudes of disparity response (difference between responses to the UCS on disparity and nondisparity trials) as a function of the number of reinforced trials.

On the question of "How does the organism tell the sanctioning properties of the afferentation?" Anokhin brings in the concept of "intention." To be effective reverse afferentation must correspond to our intention, that is, it must correspond to some provisional complex of excitations which arose before the reflex itself was formed. At the time of the arrival of the reverse afferent impulses from the UCS to the cortex the behavior will be stable only to the extent that the UCS excitation corresponds accurately to the preparatory afferent excitation which was

TABLE 1

COMPARISON OF GSR RECOVERY QUOTIENTS TO VARIOUS UCSs BEFORE AND AFTER DISPARITY EXPERIENCE ($N = 18$)

	Nondisparity UCS	Disparity UCS
Before Disparity Experience	.72	.63
After Disparity Experience	.51	.63
t ratio	2.14	
P	.05	

brought to life by the CS several seconds before the reinforcement. Put another way, the afferent apparatus produces a conclusive evaluation of the adequacy or inadequacy of the reinforcement which follows the signal. The physiological basis for this process is conceptualized as a cortical "acceptor" apparatus.

An experiment cited by Anokhin emphasizes the closeness of his work to that being reported from the USC laboratory. Two secretory CRs were established, one a tone "la" with reinforcement with rye rusks on the right side of the bench and the other a tone "fa" with reinforcement with rye rusks on the left side. During the acquisition phase no erroneous motor reactions were made—the animal went right after "la" and left after "fa." Then, on a test day dry meat was placed with the rusk on the left side. The result was a brief investigative reaction and rapid ingestion of the meat, after which the previous motor patterns became disrupted. Regardless of which CS was given the animal went to the left, suggesting that the eating of the meat led to a dominance of the reaction to the left. Further observation with rusk-only reinforcements showed refusal to eat the rusk and marked changes in the animal's behavior, suggesting what Anokhin terms a "clear-cut neurotic state."

The emphasis in Anokhin's experiment would appear to be on the disruptive effect of disparity situations rather than the basic relation between simple reinforcement concepts and the concept of sanctioning reverse afferentation. However, if one considers the matter closely, there are numerous implications for extending the perceptual disparity research described in this paper. At least one example will be considered. From Anokhin's experiment one can assume that disparity situations produce marked change in the course of the "reinforcement" process. In behavioral terms the occurrence of a disparity trial would alert the organism, lead him to expect a change in reinforcement relations. It could be argued that even the absence of the UCS, as in an acquisition test trial, may create a disparity-like situation. The result would be some effect upon the acquisition process.

Possibilities like the above can be explored by further analyzing data from experiments already described. For example, in the study discussed in connection with Figure 1, the disparity test situation contained two series, each with disparity and nondisparity trials. If the occurrence of one disparity trial has a disruptive effect, there should be a difference in responses to UCSs on nondisparity trials after a disparity experience and responses to the same UCSs prior to disparity experience. A response variable which shows this type of disruptive effect readily is the *recovery* quotient first introduced by Freeman (1948). It is an expression of the percentage return of the skin conductance to its original level occurring within a standard latency interval. If the organism's behavior is disrupted, the return to base level should be slower than when the trials are proceeding smoothly. Expressed in terms of this experimental situation one would predict a slower recovery from response to the usual UCS after a disparity experience than before. That this is the case can be seen from Table 1.

A similar effect might be expected with magnitude of response measures, that is, a larger GSR would be predicted to a given UCS after a disparity trial than before. In the experiment referred to above this did occur, response magnitude changing from 1.4 micromhos before the disparity trial to 1.7 micromhos after. A related prediction that could be tested readily from numerous published studies is that, if an interspersed acquisition test trial has disparity properties (as it is proposed here it may have), the response to the first UCS after the test trial should be different from the response to the last UCS prior to the test trial. With GSR the magnitudes should be as indicated, i.e., larger UCR after disparity.

These later considerations were introduced only to suggest some of the extended applications the disparity concept may have for explaining certain classical conditioning results. Other relations are being studied currently in order to evaluate the utility of the concept. These extend from attempts to incorporate disparity notions in a rational model of GSR behavior (Dossett & Grings, 1959) to the use of the perceptual disparity response as a measure of individual differences.

SUMMARY

A set of operations for inferring the existence of a preparatory set variable in human autonomic conditioning has been introduced and described. Termed a *perceptual disparity response* it is defined as the difference in magnitude of response between situations where receipt of stimulation following a signal cue is in accord with past experience and where receipt of stimulation is not in accord with past experience with the particular cue. Experimental data demonstrating the disparity variable are presented.

The assumption that the perceptual disparity response would be related to meaningful determining conditions in human behavior was evaluated. It was found that the disparity response magnitude was a function of the number and proportion of reinforcement trials and of the amount and direction of the stimulus change involved.

In an attempt to extend the application of the disparity concept in human conditioning, two further situations were discussed. One concerned the relation between magnitude of disparity response as a function of number of training trials, and compared this to the common learning function relating anticipatory CRs and number of training trials. The second situation evaluated the disruptive effect of a disparity experience in a sequence of regular reinforcement trials.

Brief reference was made to similar concepts and experimental work being carried out at the present time in Russia.

REFERENCES

ANOKHIN, P. K. Characteristics of the afferent apparatus of a conditioned reflex and its importance for psychology. *Voprosy Psikhologii,* 1955, No. 6, 16–38. (Reprinted, Trans. from Russian: The central nervous system and human behavior. U. S. Dept. Health, Educ., & Welfare, 1959. Pp. 241–272.)

COOK, S. W., & HARRIS, R. E. The verbal conditioning of the galvanic skin reflex. *J. exp. Psychol.,* 1937, **21**, 202–210.

CORNBECKER, J., WELCH, L., & FISICHELLI, V. Conditioning factors underlying hypnosis. *J. abnorm. soc. Psychol.,* 1949, **44**, 359–364.

DOSSETT, W. F., & GRINGS, W. W. A mathematical model for emotional arousal patterns. *Amer. Psychologist,* 1959, **14**, 425.

FREEMAN, G. L. *Energetics of behavior.* Ithaca, N. Y.: Cornell Univer. Press, 1948.

GRANT, D. A., & SHILLER, J. J. Generalization of the conditioned galvanic skin response to visual stimuli. *J. exp. Psychol.*, 1953, **46**, 309–313.

GRINGS, W. W., DOSSETT, W. F., & HONNARD, R. R. Conditioned GSR in perceptual disparity situations. *Amer. Psychologist,* 1959, **14**, 393.

GRINGS, W. W., & KIMMEL, H. D. Magnitude of GSR to expected and unexpected stimuli. Paper read at Western Psychological Association, Monterey, California, 1958.

GRINGS, W. W., & KIMMEL, H. D. Compound stimulus transfer for different sense modalities. *Psychol. Rep.*, 1959, **5**, 253–260.

GRINGS, W. W., & SHMELEV, V. Changes in GSR to a single stimulus as a result of training on a compound stimulus. *J. exp. Psychol.*, 1959, **58**, 129–133.

HILGARD, E. R., & MARQUIS, D. G. *Conditioning and learning.* New York: Appleton-Century, 1940.

HULL, C. L. Knowledge and purpose as habit mechanisms. *Psychol. Rev.*, 1930, **37**, 511–525.

KIMMEL, H. D. Amount of conditioning and intensity of conditioned stimulus. *J. exp. Psychol.*, 1959, **58**, 283–288.

KIMMEL, H. D. The relationship between direction and amount of stimulus change and amount of perceptual disparity response. *J. exp. Psychol.*, 1960, **59**, 68–72.

LACEY, J. I., & SMITH, R. L. Conditioning and generalization of unconscious anxiety. *Science,* 1954, **120**, 1045–1052.

LACEY, J. I., SMITH, R. L., & GREEN, A. Use of conditioned autonomic responses in the study of anxiety. *Psychosom. Med.*, 1955, **17**, 208–217.

MACCORQUODALE, K., & MEEHL, P. E. A reformulation of the theory of E. C. Tolman. In W. K. Estes (Ed.), *Modern Learning Theory.* New York: Appleton-Century-Crofts, 1954. Pp. 177–266.

PAVLOV, I. P. *Conditioned reflexes.* London: Oxford Univer. Press, 1927.

RAZRAN, G. H. S. Conditioning and perception. *Psychol. Rev.*, 1955, **62**, 83–95.

(Received October 20, 1958)

Psychological Review
1960, Vol. 67, No. 4, 253–268

THE PLEASURES OF SENSATION[1]

CARL PFAFFMANN

Brown University

That the senses are the channels of communication between the external environment and the nervous system is so much a truism that, except for those who espouse ESP, the senses are held to be the sole mediators of transactions between the organism and its environment. But students of the senses tend to emphasize primarily the nature of their information handling function, their discriminative capacities and limits as well as the physiological processes upon which they are based. Behavior theorists have given heed to such function in another context when they speak of the cue function of stimuli. According to these conceptions discriminative stimuli serve to steer behavior, to guide it or set the stage for the response. But in this paper I do not plan to elaborate upon such processes. Rather I wish to call attention to certain other aspects of sensory function.

In recent years, neurophysiologists have emphasized that sensory systems may mediate important functions other than discrimination. Attention has been focused on such generalized effects as arousal or alerting of the nervous system as a whole. Bremer (1935) first showed that a state of cerebral somnolence could be produced in experimental animals in which the cerebrum had been surgically isolated from the brainstem and its incoming afferent pathways. He attributed this to the interruption of sensory influx itself. Later analysis (Magoun, 1958) showed that the somnolence resulted largely from interference with the centripetal influence exerted by the reticular system, the ascending reticular activating or arousal system. Much subsequent research has shown that this system controls not only the state of arousal of the "higher centers" of the nervous system, but that it can also modulate and influence activity within the classical sensory pathways themselves (Galambos, 1956; Hagbarth & Kerr, 1954; Hernandez-Peon, 1955). The reticular system may also be significant for a variety of other psychological processes (Lindsley, 1951; Schlosberg, 1954).

Most sense organs were found to have at least two central neural pathways. One, the well-known primary projection pathway via the thalamic sensory relays to the cerebral cortex, and the second, a nonspecific pathway by way of the reticular activating system with diffuse projections to the cortex and other neural structures. The reticular system is multimodal and intramodal, for it receives inputs from many modalities. The classical projection systems are specific to one modality and may be said to mediate primarily the cognitive or discriminatory sensory functions, the nonspecific system to mediate physiological and behavioral arousal (Magoun, 1958).

Psychologists especially have emphasized that stimuli may have still other functions, especially that of reinforcement (Hebb, 1958b; Keller & Schoenfeld, 1950; Skinner, 1953; Spence, 1956). This concept has most often been discussed in purely behavioral terms until the more recent studies of direct reinforcement by intracranial stimulation (Brady, 1958; Miller, 1957; Olds, 1958; Olds & Milner, 1954). In the ensuing discussion I shall limit myself to the so-called *primary reinforcement* for there is almost no limit to the range of previously neutral stimuli that, by one method or another, can be made to acquire reinforcing properties. That *primary aversive reinforcement* has been a function of stimuli has long been known, particularly for those stimuli of high intensity which elicit defense reactions or "reflexes." Certain of these seem qualitatively more prepotent as, for example, the pain of electric shock; "rat runners" have frequently used shock as a primary negative reinforcer in their mazes or lever boxes. Some theorists hold that the pain reduction by shock termination provides the prototype of all reinforcement in the form of drive or tension reduction (Miller, 1959). The study of stimuli as *positive primary reinforcers* has, until recently, been restricted to those stimuli which are naturally related to such biologically functional activities as eating, drinking, or sexual activity. These stimuli could often be assigned a role in processes that mediated the satisfaction of a need or "homeostatic drive." Thus they were said to be related in some way to the primary biological drives. But there has been increasing evidence of late that sensory stimulation, divorced from its need or drive reducing concomitants, may function as a reinforcer in its own right. "Exteroceptive motivation" sometimes called by such names as curiosity or stimulus change (Harlow, 1953; Hebb, 1958a; Kling, Horowitz, & Delhagen, 1956; Montgomery, 1952) has been demonstrated in a number of situations. And one theorist (Hebb, 1958a, pp. 451–467, 1958b) has attempted to link such reinforcement to the reticular activating system on the grounds that changes in level of activation are reinforcing, per se, depending upon the prevailing level at the time of stimulation. Thus, in sensory isolation experiments, "the subject experienced great swings of motivation, which alternated between periods of apathy and intense desire to get back to a normal environment." Hebb concludes, "clearly man's motivation is a function of his exteroceptive stimulation." Thus, although the reinforcing function of stimuli has long been recognized and is the stock in trade of many experimentalists, there is still considerable debate as to the basis for this effect. Does it depend upon some secondary effect as need reduction, drive reduction, or change in arousal patterns in the reticular activating system? Or can we attribute these effects to sensory stimulation directly?

---

[1] Slightly revised version of Presidential Address presented at the Annual Meeting of the Eastern Psychological Association, Atlantic City, April 13, 1959.

In this paper I shall discuss the proposition that sensory stimulation per se together with its ensuing central neural events be considered as a prime determinant of reinforcement. In this context I shall discuss our experiments on the sense of taste with emphasis on the sensory physiology of taste mediated motivation as revealed in the correlation between gustatory nerve discharges and food preferences in animals. This will lead into a consideration of the relation between hedonic processes and afferent nerve discharges, preference behavior, and taste reinforcement. Finally the relation between affective sensory processes and discriminative functions will be discussed in the light of some speculations on the physiological bases for these two aspects of sensory function and their significance in behavior.

TASTE AS A MODEL SYSTEM

Although I began my studies of the sense of taste with a traditional sensory emphasis, it became increasingly apparent that the gustatory sense has certain unique features for the further

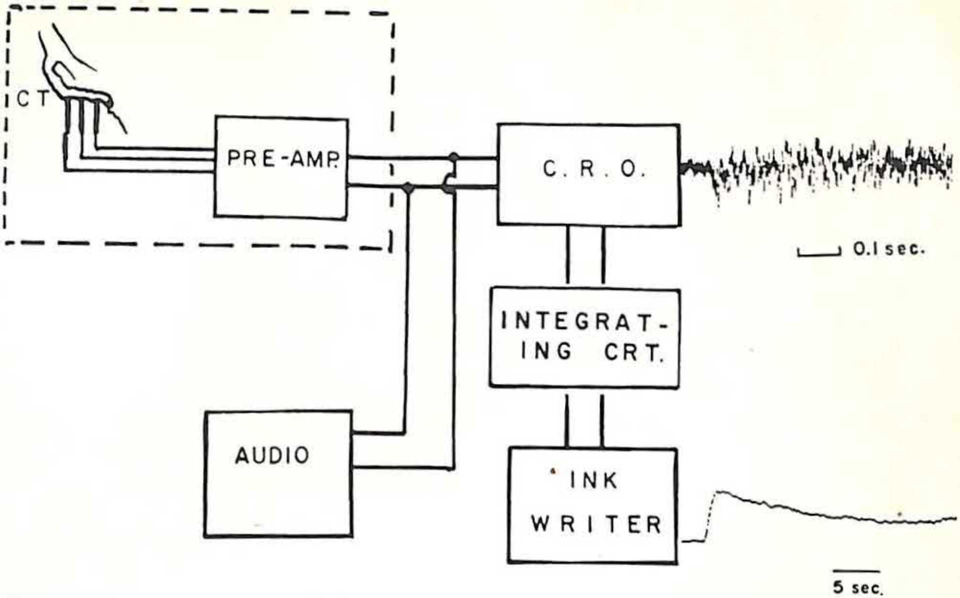


FIG. 1. A block diagram of the recording apparatus showing two types of record. The upper trace shows a typical asynchronous, multifiber discharge from a large number of nerve fibers; the lower trace shows how such activity appears when processed through the integrator. (Reproduced from the *American Journal of Clinical Nutrition.*)

understanding of behavior in general. Not only does the sense of taste possess selective receptor sensitivity that permits the discrimination of different taste stimuli, but these stimuli will elicit a number of specific consummatory responses: drinking and eating or, on the other hand, rejection. It is a relatively simple matter to demonstrate the control of behavior that taste stimuli can manifest by giving the animal a choice of water and taste solutions as Richter (1942) showed in his classic studies of self-selection. In addition, these stimuli can act as powerful reinforcers of instrumental responses leading to ingestion. Thus one and the same sense modality possesses easily demonstrable discriminative and reinforcing functions. In this respect it is an ideal model sensory-behavioral system in which to examine the relation between sensory stimulation and the reinforcing mechanisms in behavior as well as to permit the manipulation of the state of the organism by certain deprivation operations. I should, therefore, like to review briefly some of the methods we have employed and some of our experiments on taste,[2] to show how we look at the sensory continuum and its afferent neural input to the CNS which both motivates and directs behavior. In so doing I shall make use of the behavioral data of other workers, as well as the results of our own physiological and behavioral experiments. Our behavioral studies to date have utilized primarily the preferential ingestion method.

The Gustatory Afferent Discharge and Preference

Afferent nerves, like all nerves of the body, carry a series of electrical pulses which are the signs of impulse traffic up the sensory nerve. These can be

---

[2] These experiments have been supported in part by projects and grants from the Office of Naval Research, National Science Foundation, and the General Foods Corporation.

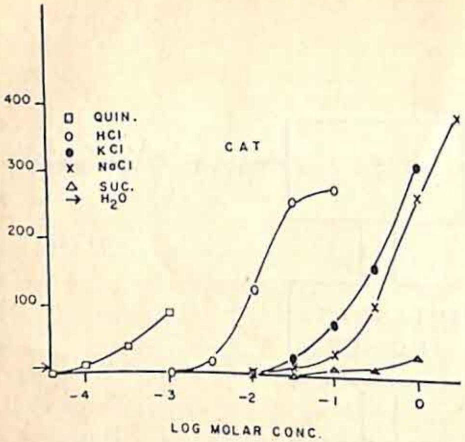


FIG. 2. Curves of taste responses in the cat to four different taste stimuli as indicated by the integrated response method. (Reproduced from the *Journal of Neurophysiology.*)

recorded by appropriate electronic devices as shown in Figure 1 (cf. Pfaffmann, 1959). I shall present only data for the over-all activity in the taste nerve as afforded by the integrating circuit illustrated in the lower trace in Figure 1. The gustatory response curves for the cat are shown in Figure 2 in terms of the magnitude of the electrical signal generated in the nerve when each of the basic taste stimuli is applied to the tongue receptor surface. By such curves we can map out the taste sensitivity of different animals (Pfaffmann, 1955). These curves do not tell us about "qualitative" differences in the nerve response. For that, analysis of the activity of single nerve fibers in the nerve is required (Pfaffmann, 1959).

From the point of view of behavior, the basic response to taste solutions is either one of acceptance or rejection so that the classical manifold of four tastes, salt, sour, bitter, and sweet, may be reduced to two behavioral classes: acceptance and rejection. We have used a typical two-bottle Richter-type preference situation to study such behavior in different species (Bare, 1949; Carpenter, 1956; Pfaffmann, 1957). Certain substances may be accepted at low, but rejected at the higher concentrations. This is so for NaCl in the rat where the well-known preference-aversion response can be demonstrated to be a function of concentration as shown in Figure 3. The solid line shows relative preference behavior

$$\frac{\text{cc. intake taste sol.}}{\text{cc. intake taste} + \text{cc. intake } H_2O}$$

Most acceptable solutions including sugar will follow this pattern when ingestion measures are used. Figure 3 also shows the neural response in the rat chorda tympani nerve to the same salt solutions. The broken line shows the integrated neural responses as a percentage of the maximum response observed. Clearly the preference behavior (i.e., response greater than 50%) is not apparent until some substantial neural input has been achieved (20%–30% of maximal discharge). The peak preference occurs at about 60% of maximal input, and the aversion rapidly sets in while the afferent discharge is still climbing. There is no change in the afferent input signal that

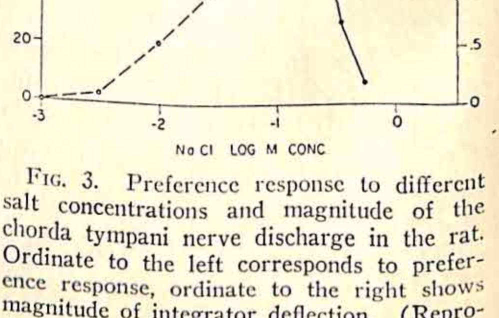


FIG. 3. Preference response to different salt concentrations and magnitude of the chorda tympani nerve discharge in the rat. Ordinate to the left corresponds to preference response, ordinate to the right shows magnitude of integrator deflection. (Reproduced from Pfaffmann, 1958.)

corresponds to the behavioral inversion point—at least we have not been able to detect any such change. Up to the inversion point, the behavior parallels the afferent discharge curve as if the positive approach to the stimulus were due directly to the afferent input. The inversion and the subsequent aversion suggest the intrusion of a secondary "stop" system. The fact that the inversion point is close to the isotonic point for sodium chloride solutions (.9% NaCl [.15M] is isosmotic with the body fluids) suggests that some postingestion or metabolic factor related to water balance has become operative.

Indeed Stellar, Hyman, and Samet (1954) have shown that the salinity of the gastric contents does affect the salt preference. Strong salt solutions intubated directly into the stomach depress the drinking of the more concentrated salt solutions. On the other hand, this is not the only stop factor for, in the same study, animals with an esophageal fistula showed the preference for hypotonic solutions and aversions for hypertonic ones even though the solutions never reached the stomach. Taste factors alone appear capable of eliciting the typical salt preference–aversion function. Thus the stop may indicate a change in "sign" of the afferent input merely as a consequence of its increasing intensity. Perhaps there is a central-neural switching when the intensity of the afferent salt discharge reaches a critical value. Most intense sensory stimuli have aversive effects.

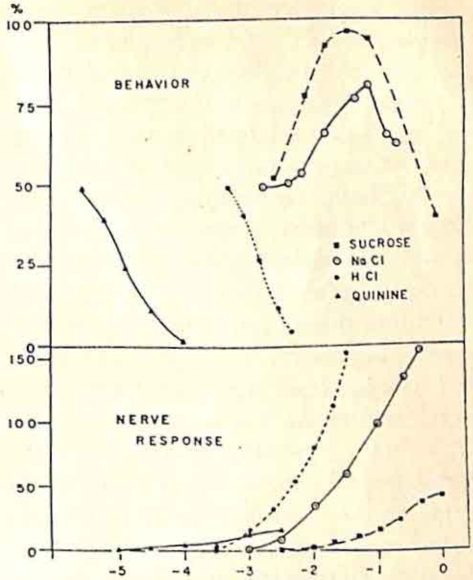


FIG. 4. Composite graph of behavioral and electrophysiological responses in the rat. Upper graphs show the percentage preference (or aversion) as a function of stimulus concentration. The integrator responses in the chorda tympani nerve to the same stimuli are shown in arbitrary units in the lower graph.

The relation of behavior to sensory afferent discharge for some of the other basic taste stimuli is shown in Figure 4. The upper figure shows the preference response curves for sugar, acid, and quinine as well as salt. We show here only the intakes above 50% for salt and sugar, remembering that at higher concentrations, the curves fall below the 50% value. The responses for acid and quinine are both aversions; no preference is shown. Note in the lower curves that the absolute magnitudes of neural response are quite different for the electrolytes as compared with the nonelectrolytes. The behavioral response to quinine is quite definite and appears before there is a clear signal "greater than the noise" for the nerve response. Here the behavioral indicator is the more sensitive. In an analogous way, the response to sucrose behaviorally is clear and definite, yet the neural response is disappointingly small. Intrinsic differences in the magnitude of the neural response, due in part to fiber size, etc., mean that sheer size of electrical signal itself cannot be correlated with behavioral effect. On the other hand, the range of effective stimulus concentration values in the two sets of determinations for each stimulus shows good agreement. The exact form or position of the preference aversion curves along the abscissa may be shifted depending upon such factors as order of presentation of stimuli, degree of hunger or physiological need, and amount of experience with the taste stimuli. It is the general nature of the relations between sensory input and preference that is shown in Figure 4.

The behavioral response to sugar solutions has been studied rather extensively, and it now seems possible to give an account of the start and stop factors in relation to sensory input. The preference curve just shown is based on ingestion of sugar over a 24-hour or longer period. Ingestion is a rising function of concentration to a peak preference when the typical inflection occurs and intake declines. This relation is found also in the single stimulus method. McCleary (1953) was one of the first to show that the intake of sugar solution is limited by an osmotic postingestion factor. Others have shown similar effects of gastric factors by studying the effects of intubation upon intake (LeMagnen, 1955; Miller, 1957; Smith & Duffy, 1957; Stellar, Hyman, & Samet, 1954). The relative frequency of choice in the brief exposure preference method, however, does not show the inflection (Young & Greene, 1953). Here there appears to be a linear relation between the level of acceptability and logarithm of concentration. Guttman's (1953) study of the rate of bar pressing as a function of concentration of sucrose solution used as the reinforcement showed that, on a continuous schedule, rate increased for the weaker concentrations but showed an inversion at the higher value. On a periodic schedule, however, the rate of bar press was found to be linear with the log of the concentration. The latter schedule provides relatively little drinking per response. But bar press rate also can be depressed by intragastric injections (Smith & Duffy, 1955).

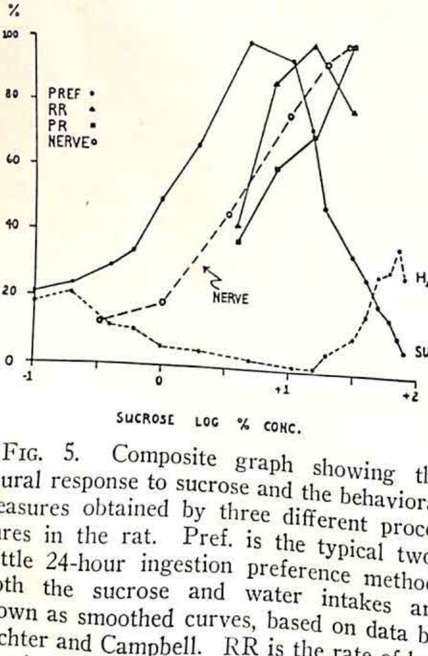


FIG. 5. Composite graph showing the neural response to sucrose and the behavioral measures obtained by three different procedures in the rat. Pref. is the typical two-bottle 24-hour ingestion preference method. Both the sucrose and water intakes are shown as smoothed curves, based on data by Richter and Campbell. RR is the rate of bar pressing during a test period for a small drop of sucrose on a regular reinforcement schedule, and PR is the same on a periodic reinforcement schedule, both based on Guttman (1953). Ordinates have been adjusted to make the maximum values in the original curves equal to 100.

The relations between bar pressing on the two schedules and the preference ingestion data of Richter and Campbell (1940) are shown in Figure 5 together with the electrophysiological response curve for the rat (Hagstrom & Pfaffmann, 1959). Note that the inversion point in the two-bottle preference test is close to the inversion point of the continuous reinforcement schedule. Both points lie close to the top of the electrophysiological sensory function. On the periodic schedule, rate rises as the sensory function increases. The fact that there is no inversion where the amount ingested is small appears to implicate the postingestion factor as a primary stop mechanism (Collier & Siskel,

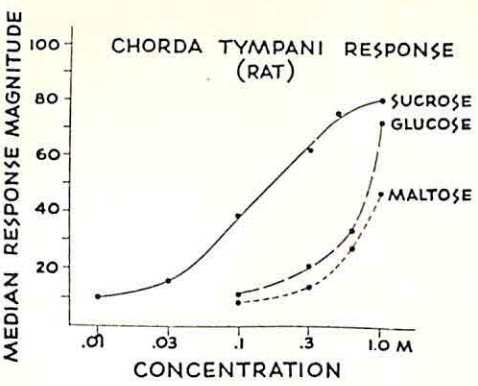


FIG. 6. Integrator response magnitudes to three sugars for the rat chorda tympani nerve. Ordinate is relative magnitude, 100 equals the response to 0.01 M NaCl. (Reproduced from *Journal of Comparative and Physiological Psychology.*)

1959; Shuford, 1959). Such a formulation would agree with Young's (1959) statement "that sugar solutions are basically hedonically positive for the rat as compared with salt which is first positive but then negative."

Studies which compared the effect of two different sugars, glucose and sucrose, on consummatory behavior, choice response, and rate of bar press likewise point to the importance of postingestive factors. Isohedonic concentrations, equally accepted in short comparative tests, were not consumed in equal amount when presented singly for longer periods. The glucose solutions were consumed in less volume than their isohedonically matched sucrose solutions, and the cumulative mean intake of both sugars was linearly related to osmotic pressure for the longer periods (Shuford, 1959). In the Skinner box the concentration of these two sugars which give equal rate of response on an aperiodic schedule (i.e., have equal reinforcing value) corresponds to the isohedonic concentrations (Shuford, 1959; Young, 1957). Guttman (1954) previously showed that the equal reinforcing solutions correspond to the equally sweet concentrations found in psychophysical experiments (Cameron, 1947; MacLeod, 1952).

Hagstrom and I (1959) have also compared the relative efficacy of sucrose and glucose as gustatory stimuli for the rat, using the electrophysiological recording method. This could be done even though the magnitude of response to sugar is relatively small. Figure 6 shows the relative responses to the two sugars, and Figure 7 shows the comparison of the equi-effective concentrations of the two sugars as determined in the bar pressing experiment, in the preference choice test, and by the electrophysiological measure. All measures agree that sucrose is more effective than glucose of equimolar concentration, although the quantitative relation is not precise, for the electrophysiological data is curvilinear, the behavioral data, linear. However, it should be remembered that the physiological data reflect only the chorda tympani response and not the taste receptors at the back of the mouth. Perhaps the curvilinear relation between the behavioral and physiological measures arises from this.

Where ingestion is minimal, behavior

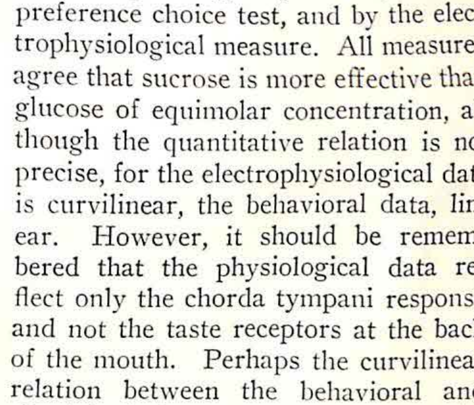


FIG. 7. Equal response concentrations of glucose and sucrose for electrophysiological measures, equal bar press rates, equi-sweet solutions and isohedonic solutions. (Reproduced from *Journal of Comparative and Physiological Psychology.*)

and sensory effectiveness seem to go together. Actually the postingestion factor seems to be especially significant in preferences based on ingestion. The glucose preference threshold is lower than that for sucrose and the intake of glucose greater than that predicted by the sensory measures. Soulairac (1947) has shown that the relative preference for sugars in a typical two-bottle preference test is correlated with their relative rates of absorption from the gastrointestinal tract.

Experiments utilizing other learning situations have been concerned with the action of sugar or saccharin as reinforcers (Collier & Siskel, 1959; Hughes, 1957; Sheffield & Roby, 1950; Sheffield, Roby, & Campbell, 1954; Smith & Duffy, 1957; Young & Shuford, 1954, 1955). There is ample evidence that the sweet taste is rewarding whether it is nutritive or nonnutritive. Performance here too is an increasing function of stimulus concentration where conditions appear to maximize the sensory effects and minimize the postingestion factor. When amount or temporal factor is such that the absolute amount or volume of stimulus or nutrient ingested per unit time is low, then purely stimulus factors seem best to account for the results. When the amount consumed per unit is large, either because of a large amount of reinforcing agent or short time between presentations, then the secondary postingestive factors intrude. As Collier and Siskel (1959) point out, it should be possible with the proper combination of concentration, volume, interval between reinforcement, and number of reinforcements to obtain a monotonic stimulus function, a nonmonotonic function with inversion points at different volumes or concentrations, or even no relation at all between amount of reinforcement and performance.

These considerations presumably apply to Campbell's (1958) measurements of the JNR, the just noticeable reinforcing difference for sucrose solutions. He determined the minimal concentration increment necessary to produce a 75% preference for the stronger of a pair of sucrose solutions in a two-bottle preference situation. Stronger solutions required a larger concentration increment than did the weak ones, and the plot of JNR against stimulus concentration showed a U-shaped function reminiscent of the more familiar Weber intensity discrimination functions. The resemblance may be due only in part to the properties of the sensory input because, as we have just shown, preference is the composite of sensory as well as postingestion factors. The osmotic effects of the stronger sugar solutions might be particularly strong. This situation contrasts with the JNR functions for the noxious stimuli where the stimulus itself may be more directly the source of the reinforcement uncomplicated by secondary factors (Campbell, 1958).

Thus we see that sensory stimulation of the mouth receptors, especially of taste, has a direct relevance for the control of ingestive behavior and the instrumental responses which lead to ingestion. Although a number of investigators have provided evidence that the mouth receptors may be bypassed and that learning can still take place, all such evidence shows that such learning is not as effective as when stimulation of the mouth receptors is included (Coppock & Chambers, 1954); Miller, 1957). Nearly all workers, whatever their theoretical predilections, have shown the importance of stimulation and "sensory contact," some would say "sensory satisfaction" (Smith & Duffy, 1957) in the reinforcement process. Indeed as Bindra (1959) concludes in his recent monograph on motivation: "Whatever the interpretation, it seems clear that, up to a certain point, an increase in sensory stimulation is a positive reinforcer" (p. 134).

The fact that certain taste stimuli control ingestion directly appears to be biologically determined. Frings (1946) has pointed out that nearly all organisms accept sugar solutions. There are exceptions perhaps related to certain aspects of metabolic or other biochemical divergences among the species. Cats cannot taste or discriminate sugar solutions (Carpenter, 1956; Pfaffmann, 1955), and birds (Pick & Kare, 1959) do not appear to show strong sucrose preferences; but these examples are remarkable largely for their divergence from what otherwise appears to be a general rule. Further there is no convincing evidence that the "sweet tooth," where it does exist, depends upon the concomitant nourishment. The drinking of nonnutritive saccharin solutions under prolonged exposure to them shows no sign of extinction such as might be expected if the preference for saccharin or "sweet" were acquired by past association with nourishment (Sheffield & Roby, 1950). Here we have "sweet for sweet's sake."

On the other side of the coin, we have been able to show that a "bitter" aversive stimulus can be made more acceptable *only temporarily* when it is paired with the alleviation of thirst in early infancy (Warren & Pfaffmann, 1958). Newborn guinea pigs were raised on a normally avoided solution of sucrose octaacetate (SOA) as the only source of water for a three-week period. At an older age these organisms showed the usual rejection of SOA. Gustatory stimuli, therefore, appear to be biologically determined as the instigator of consummatory or avoidance responses and as primary positive or negative reinforcing stimuli.

So far I have discussed the mechanisms of a variety of behaviors which appear to be under the control of sensory stimulation qua "sensory stimulation." I have not discussed the question of the affective or hedonic aspects of sensory stimulation, the pleasure of sensation, so to speak. To do so, I should now like to turn to other kinds of data derived from studies of man.

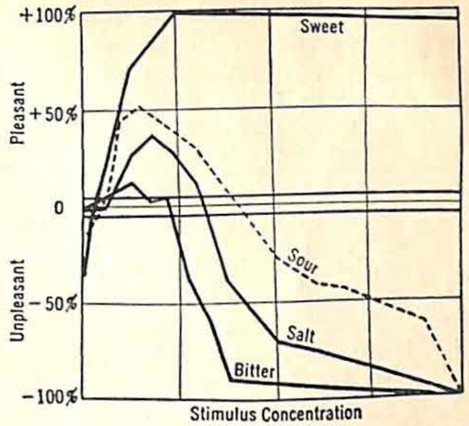


FIG. 8. The preponderance of "pleasant" or "unpleasant" judgments in relation to the concentration of taste solution. Ordinate gives percentage "pleasant" minus percentage "unpleasant." The abscissa is proportional to concentration, the full length of the base line standing for 40% cane sugar, 1.12% tartaric acid, 10% NaCl, and 0.004% quinine sulphate (by weight). (Reproduced from Pfaffmann, 1958.)

Hedonic Aspects of Sensory Stimulation

Of all the applied psychophysical fields, none have made greater use of the affective or hedonic rating scale methods, along with purely sensory testing procedures, than has the field of flavor technology. In a series of carefully controlled tests the Army Quartermaster Food Acceptance Laboratory (Pilgram, 1957), using a nine-point hedonic rating scale, has been able to predict, with good reliability, the actual choices of food and the acceptance of menus on the part of soldiers in the field. These ratings frankly ask such questions as the following:

like extremely—like very much—indifferent—dislike very much—etc.

In this case the hedonic ratings are not determined solely by stimulus properties of food, for other studies by the Quartermaster group have documented the important role played by familiarity and past experience. None the less, the frankly hedonic rating initiated by the sensory stimulation is a good predictor of actual acceptance and ingestion.

I have been impressed by the apparent similarity of our rat preference curves with those of hedonic value obtained many years ago by Engel (1928). In those experiments, subjects were asked to rate different intensities of taste stimuli as either pleasant, unpleasant, or indifferent. The data can be treated in a number of ways, but in Figure 8 we see the ratings of four different taste modalities expressed as percentage of pleasant ratings minus the percentage of unpleasant ratings for a group of seven observers using summed ratings of all observers. The abscissa is proportional to the concentrations adjusted for each of the different stimulus solutions. Note that sugar begins at a slightly unpleasant value and rises with concentration to reach a plateau. Sour, bitter, and salt all start from indifference, rise to a peak, and then fall off to unpleasantness. Sweet is predominantly pleasant, bitter predominantly unpleasant, with salt and sour intermediate. In Figure 9 I have plotted the animal preference curve and the hedonic rating by Engel's *S*s for sodium chloride solutions by concentration. The hedonic ratings here were computed simply as the percentage of the total ratings that are pleasant. Except for position along the abscissa, the two curves show a striking similarity. Beebe-Center (1951) earlier called attention to these same relations and compared the animal preference responses with hedonic ratings. Actually the more recent analysis of the postingestive factors that control the intake of sucrose solutions has done much to clarify the discrepancy between Engel's hedonic curve for sugar and the rat's preference for sugar. The sucrose hedonic curve does not turn down at the higher concentrations (see Figure 8). As noted earlier, rate of bar pressing on a periodic schedule and frequency of choice in the brief exposure test are likewise monotonically related to concentration. Thus hedonic rating and reinforcement bear the same relation to stimulus concentration.

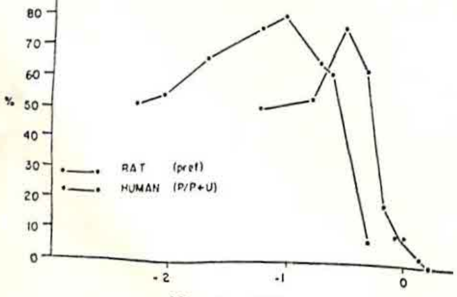


FIG. 9. Comparison of animal preference and hedonic ratings of man in response to sodium chloride solutions. (Reproduced from Pfaffmann, 1958.)

In the preference–aversion curves for other substances, the stop mechanism may be sensory (gustatory) in origin or might arise from other than postingestive effects. If the taste of strong salt solutions is aversive per se, then we might expect that, regardless of reinforcement schedule, response will fall off above the optimal salt concentration in a manner resembling the "falloff" in Engel's curves. The same might hold for saccharin which has a "bitter" sensory component at the higher concentrations. Further study of these effects is needed.

I have been emphasizing stimulus properties as prime determinants of the hedonic effect, but let me hastily point out that specific training and experience may make some unpleasant odors and tastes acceptable or even preferred. I am reminded here of the whiskey drinker's development so aptly described by Brown (1955).

> Straight whiskey, when first ingested, typically effects rather violent defense reactions. Because of this, the novice drinker usually begins with sweet liqueurs, "pink ladies," and wines, and slowly works his way through a series of beverages characterized by the gradual disappearance of cola and ginger ale additives. Finally, only plain water or even nothing need be mixed with the raw product. To the hardened drinker, straight whiskey does not taste bad—*not bad at all!* (It is thus that a product euphemistically labelled "neutral spirits" becomes indeed psychologically neutral).

Although this description appeared in defense of a drive reduction formulation of behavior and learning, the mere fact that drive reduction was supported and buttressed by the use of sweet and "pleasant" stimuli seems to have been overlooked. There is no convincing evidence that the primary relation between hedonic tone and sensory stimulation is entirely acquired. Indeed, quite the converse seems to be true.

The relation of sensory stimulation to hedonic process has been noted by earlier workers. Sherrington (1906) noted that the stimulation of contact receptors, as contrasted with distance receptors, is characterized by strong affective tone and that the contact receptors stand in very close relation to consummatory responses. Stimulation of the touch receptors of the lips and touch and taste in the mouth initiates reflex movements that precede the act of swallowing. Troland (1928) spoke of three classes of stimulus reception: nociception, beneception, and neutroception. The principal nociceptor system was pain, but other examples were hunger, thirst, the taste of bitter, strong salts and acids, and certain foul or repugnant odors, etc. In the beneceptor class were the sense organs mediating erotic behavior, the taste of sugar, and certain other odors. Vision, hearing, touch, and proprioception were relegated to neutroception. Although Troland's criterion was that of biological utility, Young (1936) pointed to the high correlation between pleasantness and beneception, unpleasantness and nociception. He also notes that there is a frequent correlation between pleasantness and approach, and unpleasantness and avoidance; and further that the affective responses to simple colors and tones are much weaker than those evoked by odors, tastes, cutaneous or organic stimuli.

In his work on palatability, Young espouses a clearly hedonistic theory of reinforcement in which he treats the affective process as a postulate, an intervening variable (Young, 1959). The reactions of animals to taste solutions then can only be described when reference is made to their positive or negative hedonic effects. Thus, sugar solutions are hedonically positive, and the level of acceptability is directly proportional to the logarithm of the concentration. Salt on the other hand does not show this one-to-one correspondence with sensory intensity. Low concentrations of salt are said to be hedonically positive for the rat, but at high concentrations are hedonically negative. Such terminology often appears simply to rename approach or avoidance behavior with hedonistic synonyms for we all know what we mean when we say in laboratory jargon that the rat likes sugar. We know he will take it in preference to water, he may press a bar for it, he may take it even when it is adulterated with something he "doesn't like" such as quinine. But such usage does not necessarily indicate the primacy of the hedonic process; we must

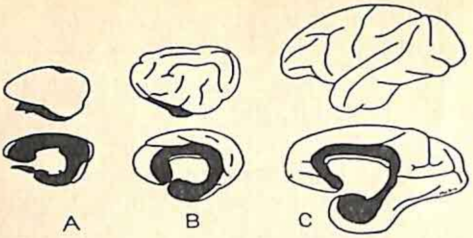


FIG. 10. Lateral and medial surfaces of brains of rabbit (A), cat (B), and monkey (C) drawn roughly to scale. The drawings illustrate that the limbic lobe, represented in black, forms a common denominator in the brains of all mammals. (Reproduced from *Journal of Neurosurgery*.)

have further experimental study of acceptance and rejection and of their relation to hedonic or affective processes.

Thus it is abundantly clear that, instigation of consummatory response (or of rejection), reinforcement of instrumental responses, and elicitation of hedonic effect are all closely related and that reproducible stimulus functions can be demonstrated for each. A definitive choice as to which of these is primary cannot yet be made. I would like to propose that sensory stimulation per se together with its ensuing central neural events be considered as a prime determinant in the chain of events culminating in acceptance behavior, reinforcement, and hedonic effect. The further study of such stimulus functions and their analysis particularly at the physiological level is a problem that should merit the highest priority.

I should like to turn to the possible physiological mechanisms in the reinforcing and hedonic functions of stimuli.

THE PHYSIOLOGY OF REINFORCEMENT AND SENSORY AFFECT

The studies of cranial self-stimulation (Brady, 1958; Olds, 1958; Olds & Milner, 1954) show that the reward and punishment systems within the brain itself, but particularly the reward systems, have been found to depend in large measure upon the limbic system. In classical neuroanatomy this was known as the rhinencephalon or smell brain, but its relation to olfaction seems less and less particular. The limbic lobe including the hippocampus completely surrounds the hilus of the hemisphere. Its various subcortical cell stations include the amygdala, septal nuclei, the hypothalamus, anterior thalamic nuclei, etc. The limbic lobe of the rabbit, cat, and monkey is shown in black in Figure 10. The limbic lobe appears to form a common denominator in the brains of all mammals.

> Its constancy of gross and microscopic structure throughout the phylogeny of the mammal contrasts strikingly with the mushrooming neopallium which surrounds it. . . . Papez theorized that the experimental evidence also points to the limbic lobe as a cortical common denominator for a variety of emotional and viscerosomatic reactions in the mammal (MacLean, 1954).

There is increasing evidence of a sensory viscerosomatic influx into the limbic system. In pathological involvements, particularly of the temporal lobe, discharges in this part of the brain may be associated with a wide variety of auras involving all the body senses as well as a great number of feeling and emotional states. Among the more purely sensory auras are crude olfactory sensations; alimentary symptoms including taste, thirst, hunger, nausea; and somatic sensations ranging from pains to paresthesia. Prolonged rhythmic responses can be elicited in the pyriform area by olfactory, gustatory, and painful stimulation (MacLean, 1954). MacLean notes that all three of these senses accent the quality and intensity of a stimulus rather than its spatial relationships.

I do not wish to give the impression that this system is essentially one for viscerosomatic sensation, for it is also implicated in a variety of complex effects such as the Klüver-Bucy syndrome (1958), memory losses (Penfield, 1958), the lowering of rage thresholds in some cases or increased docility after ablation in others (Bard & Mountcastle, 1947; Green, 1958; Schreiner & Kling, 1953; Spiegel, Miller, & Oppenheimer, 1940), and a variety of cranial self-stimulation effects (Brady, 1958; Olds, 1958; Olds & Milner, 1954). The details of these relations are beyond the scope of the present discourse. What I wish to emphasize is that there are significant sensory inputs into this system which may relate to the hedonic and reinforcing features of stimulation as compared with either the cognitive or arousal functions.

Stellar and his colleagues (Sprague, Chambers, Stellar, Liu, & Robson, 1958; Stellar, Chambers, Liu, Levitt, & Sprague 1958) at the University of Pennsylvania have made some interesting studies of chronic animals with extensive lesions in the lateral lemniscal, primary sensory pathways. I was struck with their description of the lemniscal animals, i.e., the sensorially restricted animals displayed little affect in any situation, except, perhaps the most extreme. They flex and tend to pull away from a pinch, but do not attack or show any autonomic response. They showed little or no aversive reaction to an ether cone, although they lacrimated and sneezed. Prior to the operation, they solicited petting and responded to it well; afterwards they gave no reaction to petting.

These observations suggest that the affective response to sensory stimuli might be mediated by the lemniscal and not the diffuse arousal system, which, it might be noted, already seems to have been theoretically overworked by psychologists. I would like to remind you of the "thalamic syndrome," described by Head and Holmes (1911). Lesions in the region of the thalamus are characterized by "overreaction," by excessive affectivity, for both pleasurable as well as painful sensations. This is pathology characterized by the affective modification of sensation. Although the "syndrome thalamique" is well known to clinical neurologists, its exact mechanism is still not clear. For our purposes, it is sufficient to note that in this condition, the basic hedonic responses to sensory stimulation, some of the "Pleasures of Sensation" as well as the displeasures, can be unmasked. In short, I am suggesting that the affective consequences of sensory stimulation are mediated by processes that depend upon the primary projection systems and their ramifications in thalamic and old brain neural connections. Obviously much work is required to place these speculations upon a more solid foundation.

SUMMARY

And now let me briefly summarize. I alluded briefly to the different roles of stimuli and sensory processes in the behavioral economy of the organism. I pointed out that traditionally sensory processes were studied largely in relation to discrimination or so called cognitive functions. But stimuli have been shown to have other neural and behavioral functions as arousal and reinforcement. I then went on to discuss in further detail certain of our own experiments on taste as a model system to show the relation of gustatory stimulation to the control of consummatory responses and the reinforcement of instrumental behavior. It was shown that there is increasing support for the idea that gustatory stimulation per se is capable of eliciting and reinforcing behavior in its own right. We might say "sweet for sweet's sake." I then led into the problem of hedonic or affective responses to such stimuli; and, although a good correlation between affective response and reinforcement could be demonstrated, a statement as to their exact relation must await further study, particularly at the physiological level. Finally, I speculated as to the psychophysiological mechanisms that might underlie reinforcement and the affective responses to sensory stimuli. My basic theme has been that sensory stimulation "qua stimulation" plays a significant role in the motivation as well as guidance of behavior—euphemistically we might say, in controlling behavior for the "Pleasures of Sensation."

REFERENCES

BARD, P., & MOUNTCASTLE, V. B. Some forebrain mechanisms involved in expression of rage with special reference to suppression of angry behavior. *ARNMD*, 1947, **27**, 362–404.

BARE, J. K. The specific hunger for sodium chloride in normal and adrenalectomized white rats. *J. comp. physiol. Psychol.*, 1949, **42**, 242–253.

BEEBE-CENTER, J. G. "Feeling and emotion." In H. Helson (Ed.), *Theoretical foundations of psychology.* New York: Van Nostrand, 1951. Pp. 254–317.

BINDRA, D. *Motivation, a systematic reinterpretation.* New York: Ronald, 1959.

BRADY, J. V. Temporal and emotional factors related to electrical self-stimulation of the limbic system. In, *Recticular formation of the brain.* Boston: Little, Brown, 1958. Pp. 689–704.

BREMER, F. Cerveau "isole" et physiologie due sommeil. *CR Soc. Biol., Paris,* 1935, **118**, 1235–1241.

BROWN, J. S. Pleasure-seeking behavior and the drive-reduction hypothesis. *Psychol. Rev.*, 1955, **62**, 169–179.

CAMERON, A. T. The taste sense and the relative sweetness of sugars and other substances. New York: Sugar Research Foundation, 1947. (*Sci. Rep. Ser.*, No. 9.)

CAMPBELL, B. A. Auditory and aversion thresholds of rats for bands of noise. *Science*, 1957, **125**, 596–597.

CAMPBELL, B. A. Absolute and relative sucrose preference thresholds for hungry and satiated rats. *J. comp. physiol. Psychol.*, 1958, **51**, 795–800.

CARPENTER, J. A. Species differences in taste preferences. *J. comp. physiol. Psychol.*, 1956, **49**, 139–144.

COLLIER, G., & SISKEL, M., JR. Performance as a joint function of amount of reinforcement and interreinforcement interval. *J. exp. Psychol.*, 1959, **57**, 115–120.

COPPOCK, H. W., & CHAMBERS, R. M. Reinforcement of position preference by automatic intravenous injection of glucose. *J. comp. physiol. Psychol.*, 1954, **47**, 355–357.

ENGEL, R. Experimentelle Untersuchungen über die Abhängigkeit der Lust und Unlust von der Reizstärke beim Geschmackssinn. *Arch. ges. Psychol.*, 1928, **64**, 1–36.

FRINGS, H. Biological backgrounds of the "sweet tooth." *Turtox News,* 1946, **24**, No. 8.

GALAMBOS, R. Suppression of auditory nerve activity by stimulation of efferent fibers to cochlea. *J. Neurophysiol.*, 1956, **19**, 424–437.

GREEN, J. D. The rhinencephalon and behavior. In, *Neurological basis of behavior.* Ciba Foundation Symposium. Boston: Little, Brown, 1958. Pp. 222–235.

GUTTMAN, N. Operant conditioning, extinction and periodic reinforcement in relation to concentration of sucrose used as reinforcing agent. *J. exp. Psychol.*, 1953, **46**, 213–224.

GUTTMAN, N. Equal reinforcement values for sucrose and glucose solutions compared with equal sweetness values. *J. comp. physiol. Psychol.*, 1954, **47**, 358–361.

HAGBARTH, K. E., & KERR, D. I. B. Central influences on spinal afferent conduction. *J. Neurophysiol.*, 1954, **17**, 295–307.

HAGSTROM, E. C., & PFAFFMANN, C. The relative taste effectiveness of different sugars for the rat. *J. comp. physiol. Psychol.*, 1959, **52**, 259–262.

HARLOW, H. F. Motivations as a factor in the acquisition of new responses. In, *Current theory and research in motivation.* Lincoln: Univer. Nebraska Press, 1953. Pp. 24–49.

HEAD, H., & HOLMES, G. Sensory disturbances from cerebral lesions. *Brain,* 1911, **34**, 102–254.

HEBB, D. O. Alice in wonderland or psychology among the biological sciences. In H. F. Harlow & C. N. Woolsey (Eds.), *Biological and biochemical bases of behavior.* Madison: Univer. Wisconsin Press, 1958. Pp. 451–467. (a)

HEBB, D. O. *A textbook of psychology.* Philadelphia & London: Saunders, 1958. (b)

Hernandez-Peon, R. Central mechanisms controlling conduction along central sensory pathways. *Acta. neurol. latino. Amer.*, 1955, **1**, 256–264.

Hughes, L. H. Saccharine reinforcement in a T maze. *J. comp. physiol. Psychol.*, 1957, **50**, 431–435.

Keller, F. S., & Schoenfeld, W. N. *Principles of psychology.* New York: Appleton-Century-Crofts, 1950.

Kling, J. W., Horowitz, L., & Delhagen, J. E. Light as a positive reinforcer for rat responding. *Psychol. Rep.*, 1956, **2**, 337–340.

Klüver, H. The temporal lobe syndrome produced by bilateral ablations. In, *Neurological foundations of behavior.* Ciba Foundation Symposium. Boston: Little, Brown, 1958. Pp. 175–186.

LeMagnen, J. Le role de la receptivite gustative au chlorure de sodium dans le mecanisme de regulation de la prise d'eau chez le rat blanc. *J. Physiol. Path. gen.*, 1955, **47**, 405–418.

Lindsley, D. B. Emotion. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951. Ch. 14.

McCleary, R. A. Taste and postingestion factors in specific hunger behavior. *J. comp. physiol. Psychol.*, 1953, **46**, 411–421.

MacLean, P. D. The limbic system and its hippocampal formation in animals and their possible application to man. *J. Neurosurg.*, 1954, **11**, 29–44.

MacLean, P. D., Horwitz, N. H., & Robinson, F. Olfactory-like responses in pyriform area to non-olfactory stimulation. *Yale J. Biol. Med.*, 1952, **25**, 159–172.

MacLeod, S. A construction and attempted validation of sensory sweetness scales. *J. exp. Psychol.*, 1952, **44**, 316–323.

Magoun, H. W. *The waking brain.* Springfield, Ill.: Charles C Thomas, 1958.

Miller, N. E. Experiments on motivation. *Science*, 1957, **126**, 1271–1278.

Miller, N. E. Liberalization of basic S-R concepts: Extensions to conflict behavior, motivation, and social learning. In S. Koch (Ed.), *Psychology: A study of a science.* Vol. 2. New York: McGraw-Hill, 1959.

Montgomery, K. C. Explanatory behavior and its relation to spontaneous alternation in a series of maze exposures. *J. comp. physiol. Psychol.*, 1952, **45**, 50–57.

Olds, J. Self-stimulation of the brain. *Science*, 1958, **127**, 315–324.

Olds, J. S., & Milner, P. Positive reinforcement produced by electrical stimulation of septal area and other regions of the rat brain. *J. comp. physiol. Psychol.*, 1954, **47**, 419–427.

Penfield, W. The role of the temporal cortex in recall of past experience and interpretation of the present. In, *Neurological foundations of behavior.* Ciba Foundation Symposium. Boston: Little, Brown, 1958. Pp. 149–174.

Pfaffmann, C. Gustatory nerve impulses in rat, cat, and rabbit. *J. Neurophysiol.*, 1955, **18**, 429–440.

Pfaffmann, C. Taste mechanisms in preference behavior. *Amer. J. clin. Nutr.*, 1957, **5**, 142–147.

Pfaffmann, C. *Flavor research and food acceptance.* New York: Reinhold, 1958.

Pfaffmann, C. The afferent code for sensory quality. *Amer. Psychologist*, 1959, **14**, 226–232.

Pick, H., & Kare, M. Certain aspects of taste preference in chickens and calves. Paper read at Eastern Psychological Association, Atlantic City, April, 1959.

Pilgrim, F. J. The components of food acceptance and their measurement. *Amer. J. clin. Nutr.*, 1957, **5**, 142–147.

Richter, C. P. Self-regulatory functions. *Harvey Lectures*, 1942, **38**, 63–103.

Richter, C. P., & Campbell, K. H. Taste thresholds and taste preferences of rats for five common sugars. *J. Nutr.*, 1940, **20**, 31–46.

Schlosberg, H. Three dimensions of emotion. *Psychol. Rev.*, 1954, **61**, 81–88.

Schreiner, L., & Kling, A. Behavioral change following rhinencephalic injury in the cat. *J. Neurophysiol.*, 1953, **16**, 643–659.

Sheffield, F. D., & Roby, T. B. Reward value of a nonnutritive sweet taste. *J. comp. physiol. Psychol.*, 1950, **43**, 471–481.

Sheffield, F. D., Roby, T. B., & Campbell, B. A. Drive reduction versus consummatory behavior as determinants of reinforcement. *J. comp. physiol. Psychol.*, 1954, **47**, 349–354.

Sherrington, C. *The integrative action of the nervous system.* London: Constable, 1906.

Shuford, E. H., Jr. Palatability and osmotic pressure of glucose and sucrose solutions as determinants of intake. *J. comp. physiol. Psychol.*, 1959, **52**, 150–153.

Skinner, B. F. *Science and human behavior.* New York: Macmillan, 1953.

Smith, M., & Duffy, M. The effects of intragastric injection of various substances on subsequent bar pressing. *J. comp. physiol. Psychol.*, 1955, **48**, 387–391.

SMITH, M., & DUFFY, M. Evidence for a dual reinforcing effect of sugar. *J. comp. physiol. Psychol.*, 1957, **50**, 242–247.

SOULAIRAC, A. La physiologie d'un comportement: L'Appetit glucidique et sa regulation neuro-endocrinienne chez les rongeurs. *Bull. biol.*, 1947, **81**, 273–432.

SPENCE, K. W. *Behavior theory and conditioning.* New Haven, Conn.: Yale Univer. Press, 1956.

SPIEGEL, E., MILLER, H., & OPPENHEIMER, J. Forebrain and rage reactions. *J. Neurophysiol.*, 1940, **3**, 538–548.

SPRAGUE, J. M., CHAMBERS, W. W., STELLAR, E., LIU, C. N., & ROBSON, K. Chronic reticular and lemniscal lesions in cats. *Fed. Proc.*, 1958, **17**, 154.

STELLAR, E., CHAMBERS, W. W., LIU, C. N., LEVITT, M., & SPRAGUE, J. M. Behavior of cats with chronic reticular and lemniscal lesions. *Fed. Proc.*, 1958, **17**, 156.

STELLAR, E., HYMAN, R., & SAMET, S. Gastric factors controlling water and salt solution drinking. *J. comp. physiol. Psychol.*, 1954, **47**, 220–226.

TROLAND, L. T. *The fundamentals of human motivation.* New York: Van Nostrand, 1928.

WARREN, R. P., & PFAFFMANN, C. Early experience and taste aversion. *J. comp. physiol. Psychol.*, 1958, **52**, 263–266.

YOUNG, P. T. *Motivation of behavior.* New York: Wiley, 1936.

YOUNG, P. T. Psychologic factors regulating the feeding process. *Amer. J. clin. Nutr.*, 1957, **5**, 154–161.

YOUNG, P. T. The role of affective processes in learning and motivation. *Psychol. Rev.*, 1959, **66**, 104–125.

YOUNG, P. T., & GREENE, J. T. Quantity of food ingested as a measure of relative acceptability. *J. comp. physiol. Psychol.*, 1953, **46**, 288–294.

YOUNG, P. T., & SHUFORD, E. H., JR. Intensity, duration, and repetition of hedonic processes as related to acquisition of motives. *J. comp. physiol. Psychol.*, 1954, **47**, 298–305.

YOUNG, P. T., & SHUFORD, E. H., JR. Quantitative control of motivation through sucrose solutions of different concentrations. *J. comp. physiol. Psychol.*, 1955, **48**, 114–118.

(Received October 29, 1959)

Psychological Review
1960, Vol. 67, No. 4, 269–271

THEORETICAL NOTES

THE CONCEPT OF STATISTICAL SIGNIFICANCE AND THE CONTROVERSY ABOUT ONE-TAILED TESTS

H. J. EYSENCK

University of London

Several controversial papers regarding the uses and abuses of the one-tailed test of significance have recently appeared (Burke 1953, 1954; Goldfried 1959; Hick 1953; Jones 1952, 1954; Marks 1951, 1953). As Goldfried (1959) points out, "the important question debated is not *if* it should be used, but rather *when* it should be used." It is suggested here that most of the disagreements emerging from this controversy stem from a misunderstanding of the term "significance," and it is further suggested that the same misunderstanding runs through many discussions of two-tailed tests as well. It will be suggested that in the sense in which Goldfried's statement is meant, it has been the wrong question which has been debated; neither one-tailed nor two-tailed tests should be used at all in the sense envisaged by most of the writers quoted.

The outcome of the statistical examination of experimental results is always stated in terms of the probability of disconfirmation of the null hypothesis; the set of values which these p values can take is continuous in the interval from 0 to 1. It is customary to take arbitrary p values, such as .05 and .01, and use them to dichotomize this continuum into a *significant* and an *insignificant* portion. This habit has no obvious advantage, if what is intended is merely a restatement of the probability values; these are already given in any case and are far more precise than a simple dichotomous statement. Indeed, gross absurdities result from taking these verbal statements too seriously; the difference between a C.R. of 1.90 and another of 2.00 is quite negligible, yet one falls on one side of the dichotomy, the other on the other side. This has led to such summary statements as: "almost significant," or "significant at the 10% level." If the verbal dichotomous scale is not satisfactory—as it clearly is not—the answer surely is to keep to the continuous p scale, rather than subdivide the verbal scale.

However, surplus meaning has accrued to the word "significant," and it has become a shibboleth which divides the successful from the unsuccessful research. It is frequently interpreted as almost guaranteeing reproducibility of results, while failure to reach significance is interpreted as disconfirmation. Hence the urgent desire to achieve respectability and significance by one-tailed tests, if need be, and the argument regarding when the cachet of "significance" can be bestowed upon a research result. Yet the argument, and the achievement or nonachievement of significance, do not alter the facts of the case, which are contained in the statement of the p value of the results. Anything beyond these facts depends upon interpretation, and is subjective; it does not alter the facts of the case in the slightest.

As an example of the necessity of interpretation, consider the a priori probability of the conclusion. Suppose that an experiment on ESP were carried out with all the precautions which human ingenuity can devise, so that even the most sceptical had to agree that no fault could be found with the experimental design. *Suppose* also that a p value of .05 were achieved. Would this be considered "significant," in the sense of guaranteeing reproducibility? Critics would point out quite rightly that where the a priori probability is very low, as in this case, much higher p values would be required to carry significance. Logicians are agreed that interpretation of experimental results must call on all available knowledge about the subject in question; a priori probability is a kind of summary statement of much of this knowledge. It cannot be overlooked in arriving at a conclusion regarding "significance" when the term carries the surplus meaning indicated.

That interpretation comes into the problem very much is clear when we look at such conditions as those suggested by Kimmel (1957) as criteria for the use of one-tailed tests. He suggests, for instance, that they may be used if results in the opposite direction would be psychologically meaningless or could not be deduced from any psychological theory. These are obviously not objective criteria, but depend on what the author (or reader) considers psychologically meaningless, or the kind of theory he may hold. Opinions will differ, and consequently some readers will agree to the use of the one-tailed test in a particular case, others will not. Thus to some readers the results will appear *significant*, to others *insignificant*.

The whole argument seems to be about *words*, not about *facts:* Is the word "significant" to be used in a given situation, or is it not? This would only matter if the word carried some objective meaning not contained in the probability figures; we have argued that it does carry surplus meaning, but that this is not of an objective kind. Consequently, nothing important is changed by omitting the term altogether in the report, leaving interpretation to the reader. After all, the only true proof of reproducibility is reproduction! Verbal assertions of "significance" have no more meaning than the *droit du pour* at the court of Louis XIV.

The solution is to separate quite clearly and decisively the *objective statement of the probability of disproof* of the *null hypothesis* (by means of a two-tailed test), and the *subjective evaluation and interpretation of the results.* The reader would be able to accept the first statement as a statement of fact and would then be able to judge for himself the arguments presented by the author regarding the *meaning* of these facts. These arguments might be based on results of previous experiments, predictions made on the basis of more or less widely accepted theories, number of cases involved, a priori lack of accepability of the conclusions, and other similar grounds; an explicit statement of the arguments would enable the reader to decide for himself the acceptability of the conclusions in a manner precluded by the simple statement of one-tailed probability. *A statement of one-tail probability is not a statement of fact, but of opinion, and should not be offered instead of, but only in addition to, the factual two-tailed probability;* if it is offered at all, it should be accompanied by a full statement of the arguments in favor of its facilitating a more meaningful interpretation of the data. In the writer's opinion, it would be better to drop such statements of one-tailed probability altogether and rely entirely on appropriate argumentation to establish the meaning of the observed (two-tailed) probabilities.

Implicit in this recommendation is the corollary that the mechanical evaluation of experimental results in terms of "significant" and "not significant" be dropped outright. Interpretation is implicit in the statement of one-tailed probabilities, but it is also implicit in the statement of two-tailed probabilities if these are *automatically* interpreted as being significant or not significant, with all the surplus meaning carried by these terms. The experimenter should give his (two-tailed) p values and then proceed to argue regarding the acceptability of the conclusions on the basis already indicated. There have appeared in the literature solemn discussions about the possible causes for discrepancies between two experiments, one of which gave significant, the other insignificant results; yet the respective t values were almost identical, one lying just on the one side, the other just on the other side, of the arbitrary 5% line. Such arguments are unrealistic and would be avoided if p values were compared, rather than verbal statements. Two experiments giving p values of .048

and .056 are in excellent agreement, although one is significant, while the other is not.

To summarize the main point of this note briefly, we would say that verbal statements regarding "significance" are at best supererogatory restatements in an inconvenient dichotomous form of results already properly stated in terms of a continuous system of p values; at worst they carry unjustified surplus meaning of an entirely subjective kind under the guise of an objective and mathematically meaningful statement. Subjective judgments of reproducibility cannot reasonably be based on the mechanical application of a rule of thumb whose only usefulness lies in the elementary instruction of undergraduates lacking in mathematical background; if they are to be made at all they demand complex consideration of a priori probabilities. It is suggested that the accurate and factual statement of probabilities (two-tailed) should be mandatory and that all subjective considerations, arguments, and judgments should be clearly separated from such factual statements. It is implied that judgments of "significance" belong with the subjective side, and it is also implied that the calculation of p values on the basis of one-tailed tests has no place in psychology.

REFERENCES

BURKE, C. J. A brief note on one-tailed tests. *Psychol. Bull.*, 1953, **50**, 384–387.

BURKE, C. J. Further remarks on one-tailed tests. *Psychol. Bull.*, 1954, **51**, 587–590.

GOLDFRIED, M. R. One tailed tests and "unexpected" results. *Psychol. Rev.*, 1959, **66**, 79–80.

HICK, W. E. A note on one-tailed and two-tailed tests. *Psychol. Rev.*, 1952, **59**, 316–318.

JONES, L. V. Tests of hypotheses: One-sided and two-sided alternatives. *Psychol. Bull.*, 1952, **49**, 43–46.

JONES, L. V. A rejoinder on one-tailed tests. *Psychol. Bull.*, 1954, **51**, 585–586.

KIMMEL, H. D. Three criteria for the use of one-tailed tests. *Psychol. Bull.*, 1957, **54**, 351–353.

MARKS, M. R. Two kinds of experiment distinguished in terms of statistical operations. *Psychol. Bull.*, 1951, **58**, 179–184.

MARKS, M. R. One- and two-tailed tests. *Psychol. Rev.*, 1953, **60**, 203–208.

(Received August 28, 1959)

Psychological Review
1960, Vol. 67, No. 4, 272–276

A NOTE ON THEORY AND METHODOLOGY IN THE STUDY OF FIGURAL AFTEREFFECTS

R. ALLEN GARDNER

Quartermaster Research and Engineering Center, Natick, Massachusetts

Recently, Mathews and Wertheimer have derived a theorem from the Kohler-Wallach (1944) theory of figural aftereffects in order to provide an experimental test to discriminate between that theory and the Osgood-Heyer (1952) theory. This test must fail, however, because the principle that is invoked is contained in both theories, and the same prediction is at least as readily derived from either. Instead, the contribution of Mathews and Wertheimer lies in drawing attention to some commonly overlooked, but logically necessary consequences of a satiation-type theory. The rule that displacements will be uniformly symmetrical with respect to the contours of any I-figure has come to be accepted as a general statement of Kohler's theory and empirical findings. For example, Sagara and Oyama (1957) reported several experiments in which asymmetrical displacements were found as a contradiction of Kohler and Wallach (1944). The interpretation that Kohler and Wallach gave for this same finding in their original paper is rejected by Sagara and Oyama as a special ad hoc suspension of the rule of symmetry. In part, the object of the following analysis is to show that asymmetrical displacements are a logically necessary consequence of Kohler's assumptions in a wide variety of situations, while the rule of symmetry only applies to a restricted class of figural aftereffects.

The satiation of Kohler and Wallach and the neural adaptation of Osgood and Heyer are processes which arise from the contours of a figure when it is seen against a contrasting field. *Satiation* spreads to regions about the contours of the figure in a manner determined by electrotonic field-forces assumed to exist in the visual projection area of the cortex. *Neural adaptation* spreads about the contours in a manner determined by the statistical spread of neural excitation in the visual cortex caused by "the simultaneous action of physiological nystagmus and reciprocal overlap of dendritic processes" (Osgood & Heyer, 1952, p. 102). As Osgood and Heyer freely admit, once each theory has accomplished its genuflections to the particular sort of ritualistic physiological speculation favored by its authors, the results are essentially equivalent. With time, a gradient of satiation (or neural adaptation) is formed about the contour. The gradient is highest directly over the contour and decreases with distance in either direction, tending to be sigmoid in shape due to an assumed flattening at the peak. In general, displacements will be from regions where the gradient is high to regions where it is low. The derivation up to this point is illustrated in Fig. 1, where CD represents a line presented as an I-figure, P and P′ represent the loci of T-figures presented after CD has been removed, and the arrows indicate the direction of displacements. If either P or P′ is brought too close to CD, the gradients will extend symmetrically in both directions so that no displacement can occur. The extent of displacement has been found to increase with distance from the contour up to a maximum and then decrease again (the distance paradox). To account for this, both theories assume that the extent of displacement is a complex function of the height and the steepness of the gradients. Such symmetry of direction and extent, with CD representing the axis of symmetry, can be expected only as long as the I- and T-figures are analogous to the diagram of Fig. 1 which will be referred to here as Paradigm I. The analogy will continue to hold when the I-line, CD, is widened to become an oblong or square, but only so long as the T-figure does not become appreciably smaller than the I-figure.

Paradigm I is the illustration most

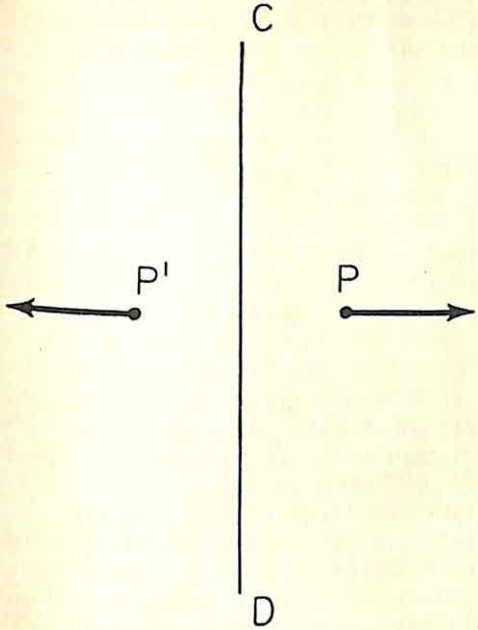


FIG. 1. Paradigm I.

frequently mentioned in the theoretical discussions of both Kohler and Wallach and Osgood and Heyer. Also, analogies of Paradigm I comprise the more intuitively obvious and striking of the various demonstrations reported by Kohler and Wallach, as well as those most reliably confirmed by subsequent investigators.[1] Thus, it is understandable that this simple paradigm has come to be accepted as the general model of figural aftereffects according to satiation theory. As Kohler would say, "Nothing could be farther from the truth." Consider the diagram in Fig. 2. Here the square ABCD represents a solid I-figure presented against a contrasting field (e.g., black square on white field). The point P represents the locus of a T-figure appreciably smaller than ABCD that is presented outside the contours of the I-figure and near the contour CD, while Q represents the locus of a similar T-figure presented near CD but within the contours of the I-figure. The solid arrows represent the symmetrical displacements on either side of CD that would be predicted merely by superimposing the I-line of Paradigm I on the new diagram. Clearly, this would not be justified. Unless the I-figure is so wide or the I-period is so short that the tissue satiating currents flowing from the opposite contour are negligible, displacement at Q should not be symmetrical with displacement at P. Without doing violence to the rest of the theory it is necessary to suppose that the satiating currents arising from the opposite contours, AB and CD, combine with each other in many cases. As a result, the satiation at loci inside the contours will always be greater than that at an equal distance outside the contours. At first, the cross sectional pattern of satiation will be saddle-shaped with the level of satiation in the cleft of the saddle rising more rapidly than that of the outward sloping gradients. With time a relatively flat dome-shaped pattern will emerge which spans the distance between the contours. Eventually, a homogeneous degree of satiation will be obtained between the contours and extending for some distance beyond them (Kohler & Wallach, 1944, p. 343–344). A precise statement of the rate at which each stage will appear and then give way to the next is beyond the scope of the theory in its present form. Nevertheless, with a suitable choice of I-figure, T-figure, and I-period, it should be possible to find a locus Q between the area of greatest satiation and the contour CD such that the direction of displacement will be *towards*

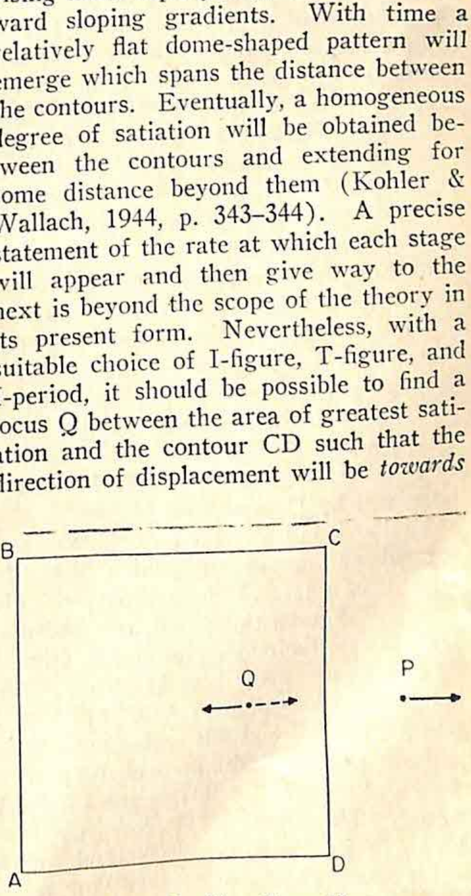


FIG. 2. Paradigm II.

---

[1] Extensive reviews of this literature have been published recently by McEwen (1958), Sagara and Oyama (1957), and Spitz (1958).

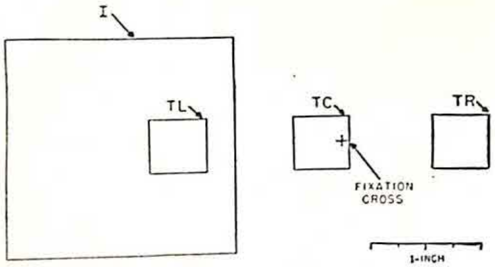


FIG. 3. I- and T-figures used by Mathews and Wertheimer (1958).

CD. This is indicated by the broken arrow in Fig. 2.

Despite the claim of Mathews and Wertheimer (1958), which they have documented with a personal communication from Osgood, precisely the same theorem can be derived from the Osgood-Heyer theory. Osgood and Heyer (1952, p. 108) explicitly state that the statistical distributions of neural adaptation arising from opposite contours must overlap and summate with each other in a manner analogous to the satiation of Kohler and Wallach. Their Fig. 2 (1952, p. 105) is a cross sectional representation of Paradigm I. It becomes Paradigm II merely by superimposing upon it the overlapping gradient of neural adaptation which might be expected from a second contour. If this is done, the predictions of the Mathews-Wertheimer theorem must follow. Of course, to obtain a prediction for a specific I-figure it will no longer be possible to avoid specifying the parameters of the statistical distributions as it was in the original analysis presented by Osgood and Heyer.

Mathews and Wertheimer (1958) performed an experiment which they designed as a test of their theorem. The I- and T-figures they used are drawn to scale from their description in Fig. 3. The I-figure was a solid, black, 2-in. square corresponding to ABCD in Fig. 2, while TL, TC, and TR were three ½-in. squares presented simultaneously as a T-figure after a 2-min. inspection of the I-figure. The locus of TL corresponds to locus Q in Paradigm II. Each of 16 observers (*O*s) was asked to report which distance was greater, that between TL and TC or that between TC and TR, and 14 chose the right hand distance as the greater. According to Mathews and Wertheimer: "This result favors the Kohler-Wallach prediction, by bimodal expansion, at the 1% level." Since no further data are submitted, we must suppose that this probability level was arrived at by assuming that 16 other *O*s who had not inspected the I-figure—that is, a control group—would have been equally divided in their choices. Unfortunately, those of us who do not concur in this admittedly common practice of assuming the performance of a control group must remain in doubt as to the outcome of the experiment. Further elaboration of this point is unnecessary as the importance of such a control in the study of figural aftereffects has already been discussed at some length by Brown (1953) and Wertheimer (1954).

In any event, even with a control group, the particular T-figure chosen by Mathews and Wertheimer could never be used to study the displacement of the square TL. All objects in the T-figure must be assumed to exhibit some displacement. Consequently, the best possible measurement of the extent or direction of displacement is the relative displacement of a near T-object when compared with that of a second, remote T-object. When the relative displacements of both the near and the remote T-objects are referred to that of an intermediate, third T-object no meaningful result can be obtained. The *O*s could report that the distance TC—TR is greater than the distance TL—TC when both TL and TC are displaced *away* from the contour in any case where the displacement of TR is greater than the combined displacement of TL and TC. Indeed, this might occur very commonly due to the distance paradox. Similarly, the distance TL—TC could be reported as greater even when TL is displaced *towards* the contour in any case where the displacement of TC away from the contour is greater than the combined displacement of TL and TR. The ubiquitous distance paradox would readily account for this outcome also.

Fortunately, Fox (1951) has already reported two technically more adequate experiments that fulfill the topological requirements of the Mathews-Wertheimer theorem. In Experiment I, Fox found that T-figures presented on the contour or just inside of it were displaced outwards, but to a smaller extent than T-figures presented outside the contour. Using somewhat different I-figures, in Experiment II he found that displacements were symmetrically away from the nearest contour for a variety of loci on either side of the contour. Fox concluded in favor of Kohler and Wallach on the basis of Experiment II, attributing the asymmetrical results of Experiment I to an artifact not present in Experiment II. It might at first appear that in order to preserve the Mathews-Wertheimer theorem, Fox should have based his conclusion on the results of Experiment I and explained away the symmetrical results of Experiment II. But this, too, would oversimplify the situation and underestimate the full power of the satiation theory. Clearly, both the initial pattern of satiation and its variation as a function of time must depend upon the particular I-figure that is in question. In the complex situation described by Paradigm II Fox's results can be accounted for simply by noting that the I-period he used must have been long enough to produce asymmetrical displacements with the I- and T-figures of Experiment I, but not with those of Experiment II.

The Mathews-Wertheimer theorem represents Case 1 of Paradigm II. It is also possible to derive its converse, Case 2, where the direction of displacement at loci outside the contours reverses with time under certain conditions. That this case, too, is implied as a consequence of satiation theory was noted by Kohler and Wallach (1944):

> Suppose that the T-squares lie half inside and half outside the area of the I-oblong [on opposite contours]. If the inspection period is not too short the distance between these squares will decrease. Even with T-squares which lie entirely outside the oblong but just adjacent to it the same will be observed—always provided that the inspection period has been sufficiently long . . . the direction of the effect depends upon the degree in which the medium has been altered by previous inspection. Also, during late stages of the inspection period a figure process may not affect the medium in quite the same way as it does in the beginning. Incidentally, in the same arrangement slight variations of the I-object may turn the effect in one direction or the other (p. 297).

The existence of Case 2 displacements towards the contour that Kohler and Wallach reported have since been confirmed by Smith (1954) and by several Japanese studies reviewed by Sagara and Oyama (1957). The explanation seems to depend upon the existence of I-figures which give rise to gradients which are steeper inside the contour than outside the contour. Both Kohler and Wallach (1944, pp. 339–340) and Osgood and Heyer (1952, p. 108) invoke this principle to account for an otherwise anomalous discrepancy between outline and solid figures. In such a case the peak will be shifted outwards somewhat, and further summation and interaction with the gradient extending from the opposite contour should increase this shift. The effect will not always be observed, of course; but with a suitable choice of I-figure, T-figure, and I-period, it should be possible to find a locus *p* outside the contour such that displacement will be towards the contour. In connection with this, Sagara and Oyama (1957) report that where Case 2 reversals were found the extent of nonreversal displacement was greater inside the contours than outside. This is precisely the result that would be expected if the inside gradient is steeper than the outside gradient in Case 2.

There is essentially no contradiction between the predictions of Case 1 and those of Case 2. Each is appropriate to a separate class of I-figures. It is intuitively obvious that different classes of I-figures must give rise to different patterns of electrical activity in the visual cortex. To claim otherwise is to do violence to language. However, the various patterns are related by a common neurological mechanism. According to Kohler and Wallach, this mechanism is the electrotonic field-forces of satiation. According to Osgood and Heyer, it is the statistical distribution of neural adaptation. Neurological evidence has been cited on both sides. To be sure, the neurological data available to the authors were not sufficient to permit either theory to specify the particular pattern of cortical activity that would be produced by any particular set of conditions. This is understandable when we consider that the evidence that is cited pertains to phenomena that take place within fractions of a second, while figural aftereffects can extend over several minutes at least. But, with disarming frankness, both Kohler and Wallach and Osgood and Heyer admit the need for much further research in this area. In the interim, the gap is bridged by boldly assuming that in cases of Paradigm I the long-term neurological processes will be analogous to the short-term processes, while in cases of Paradigm II summation and interaction over time will produce precisely the gradients that are necessary to account for the figural aftereffects that occur. Thus, whether the neurophysiologists of the future opt for field-forces, or statistically independent units, or—as in modern physics—some happy combination of the two, it is probably safe to say now that perceptual phenomena are related to neural activity.

REFERENCES

BROWN, K. T. Methodology for studying figural after-effects and practice effects in the Muller-Lyer illusion. *Amer. J. Psychol.*, 1953, **66**, 629–634.

FOX, B. H. Figural after-effects: Satiation and adaptation. *J. exp. Psychol.*, 1951, **42**, 317–326.

KOHLER, W., & WALLACH, H. Figural after-effects: An investigation of visual processes. *Proc. Amer. Phil. Soc.*, 1944, **88**, 269–357.

MCEWEN, P. Figural after-effects. *Brit. J. Psychol., Monogr. Suppl.*, 1958, **31**, 1–106.

MATHEWS, T., & WERTHEIMER, M. An experimental test of the Kohler-Wallach and the Osgood-Heyer theories of figural after-effects. *Amer. J. Psychol.*, 1958, **71**, 611–612.

OSGOOD, C. E., & HEYER, A. W., JR. A new interpretation of figural after-effects. *Psychol. Rev.*, 1952, **59**, 98–118.

SAGARA, M., & OYAMA, T. Experimental studies on figural after-effects in Japan. *Psychol. Bull.*, 1957, **54**, 327–338.

SMITH, K. R. "Attraction" in figural after-effects. *Amer. J. Psychol.*, 1954, **67**, 174–176.

SPITZ, H. H. The present status of the Kohler-Wallach theory of satiation. *Psychol. Bull.*, 1958, **55**, 1–28.

WERTHEIMER, M. Constant errors in the measurement of figural after-effects. *Amer. J. Psychol.*, 1954, **67**, 543–546.

(Received August 28, 1959)

Psychological Review
1960, Vol. 67, No. 4, 277–278

SCALE CLASSIFICATION AND STATISTICS

JOHN GAITO

Wilkes College

The classification of psychological scales into nominal, ordinal, interval, and ratio categories by Stevens (1951) has been a significant contribution to psychophysical and measurement theory. This classification has been widely accepted and applied by psychologists in conversation, teaching, and in publication (e.g., Edwards, 1958; Guilford, 1954; Senders, 1958; Siegel, 1956). In considering these scales Stevens has specified the appropriate statistical measures for use with each. This latter aspect has had a tremendous impact on writers of statistical texts. For example, these categories form the basis for the texts by Senders (1958) and by Siegel (1956). There is no doubt that this classification has great practical and theoretical value. However, we feel that the application of this classification has led to some misunderstanding in regard to the use of various statistical techniques and to an overemphasis on the utility of nonparametric techniques in psychological research.

As an example of this tendency Siegel (1956, p. 26) maintains that nonparametric tests of significance should be used with subinterval type data. He lists the requirement of interval scale as one of the assumptions for the use of the analysis of variance. However, this assumption cannot be found if one looks to the mathematical bases of assumptions (Eisenhart, 1947). In fact a noted statistician, Kempthorne (1955), after showing mathematically that the normal theory analysis of variance test can approximate the randomization test, states that

> This serves as some theoretical basis for the fact which has been noticed by most statisticians, that the level of significance of the analysis of variance test for differences between treatments is *little affected by the choice of a scale of measurement for analysis* [italics added].

The important consideration for the use of the analysis of variance is not that the data have certain scale properties but that *the data can be related to the normal distribution,* plus approximating the other assumptions of independence and homogeneity of errors. Furthermore, it must be emphasized that the normal distribution refers to the distribution of those portions of the data which are used as the appropriate estimates of error and not necessarily to the distribution of the observations. This is an important consideration because in multivariable designs the variation within each group may not distribute normally but the portions of certain interaction terms may. In many of these designs some of the interaction effects provide the valid estimate of error in testing certain effects.

It would seem that the comments concerning the relation between the various scales and the appropriate statistical procedures should serve merely as a guide. In some cases empirical data may indicate that rigid adherence to such statistical methods is not required and wasteful of the data. The author has propounded elsewhere this viewpoint and cited relevant data concerning the analysis of variance technique (Gaito, 1959).

Furthermore, the same data may be considered to have the properties of two or more scales, depending on the context in which it is considered. For example, if we look at the response of one subject (S) to a single item, the properties of the data are those of a nominal scale, i.e., right or wrong. However, if we concentrate on the total score for one S or the total scores for a group of Ss, we have at least an ordinal scale. This is similar to the situation of approximating the binomial distribution (a distribution having nominal, or possibly ordinal, scale properties) by the normal distribution (a distribution having at least ordinal, possibly interval, scale properties) when n increases and p is close to .5. Another example is the relation between the sign test and the binomial distribution test. For a given set of data these two methods will give the same result; yet Siegel lists the former as of an ordinal nature, the latter as nominal data.

It appears that in dealing with descriptive statistics the comments made about the use of nonparametric statistics for subinterval-type data are most appropriate. However, in dealing with tests of significance such statements are not as appropriate, even though descriptive statistics are involved in the procedure. For each test the mathematical requirements are expressed in the assumptions, and nothing concerning specific scale properties is required. Supplementing the requirements will be evidence to indicate the effects on the data of failure to meet the assumptions (Gaito, 1959; Lindquist, 1953). Thus if empirical research indicates little effect on the data when assumptions are not met or merely approximated, the investigator may use a parametric technique and allow for the effects of the deviations from assumptions in his interpretation.

It is encouraging to note that some individuals have been reluctant to embrace wholeheartedly the nonparametric technique (e.g., Guilford, 1954; McNemar, 1957, 1958; Savage, 1957). In a review of the statistical literature in the *Annual Review of Psychology*, Grant (1959, p. 137) appropriately summarizes this position by the comment: "*Some much-needed negative thinking has recently appeared on nonparametric techniques*" [italics added].

In conclusion we wish to restate that the statements concerning scale properties and statistical procedures should be guides and that context, mathematical assumptions of statistical procedures, and the results of research concerned with failure to satisfy assumptions should be the ultimate determiners of the choice of statistical techniques.

REFERENCES

EDWARDS, A. E. *Statistical analysis.* New York: Rinehart, 1958.

EISENHART, C. The assumptions underlying the analysis of variance. *Biometrics*, 1947, **3**, 1–21.

GAITO, J. Nonparametric methods in psychological research. *Psychol. Rep.*, 1959, **5**, 115–125.

GUILFORD, J. P. *Psychometric methods.* New York: Wiley, 1954.

KEMPTHORNE, O. The randomization theory of experimental inference. *J. Amer. Statist. Ass.*, 1955, **50**, 946–967.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* New York: Houghton Mifflin, 1953.

MCNEMAR, Q. On Wilson's distribution-free test of analysis of variance hypotheses. *Psychol. Bull.*, 1957, **54**, 361–362.

MCNEMAR, Q. More on the Wilson test. *Psychol. Bull.*, 1958, **55**, 334–335.

SAVAGE, I. R. Nonparametric statistics. *J. Amer. Statist. Ass.*, 1957, **52**, 331–344.

SENDERS, V. L. *Measurement and statistics.* New York: Oxford Univer. Press, 1958.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

STEVENS, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951.

(Received September 11, 1959)

VOL. 67, NO. 5 SEPTEMBER 1960

THE PSYCHOLOGICAL REVIEW

DISCRIMINATION AND LEARNING WITHOUT AWARENESS:

A METHODOLOGICAL SURVEY AND EVALUATION [1]

CHARLES W. ERIKSEN

University of Illinois

The importance of a concept of the unconscious in psychology does not need to be documented. Most, if not all, current theories of personality employ the concept in their treatment of such diverse problems as defense mechanisms, motivation, psychopathology, and psychotherapy. Despite the importance of this concept little or no progress has been made in refining and objectifying it. The current treatment of the unconscious is little if any more advanced than the loose animistic concept devised by Freud. This looseness gives rise to excessive surplus meaning and many powers (occasionally bordering on the mystical) are glibly and uncritically attributed to this pervasive process.

While there has been a resurgence of experimental interest in unconscious phenomena, this research for the most part tends to be primarily demonstrational in nature rather than parametric. The complexity of the experimental operations affords numerous opportunities for artifacts to occur and the results are generally overinterpreted. In this latter connection little or no attempts are made to relate the phenomena to the general body of experimentally derived psychological knowledge.

This paper presents a representative survey of experimental approaches to unconscious phenomena with the purpose of examining methodological and definitional problems that are encountered. An attempt is made to separate different phenomena that have been lumped together under the general concept of the unconscious and to treat these phenomena in terms of current laboratory derived concepts.

PROBLEMS OF DEFINITIONS

The nonscientific nature of current conceptions of the unconscious is attested by the vagueness of definitions of what constitutes conscious and unconscious behavior. Not only are the definitions vague and operationally inadequate but there are several different terms that are used, sometimes to designate the same phenomena, other times in a partially nonoverlapping way. Such terms as "consciousness," "awareness," "insight," and "attention" have all been used to describe the same phenomena. I do not intend to trace the differences or the commonalities among these different terms and their application to specific problems but merely to indicate the nature of the confusion that exists. In the remainder of the paper I shall use the terms awareness and unawareness to designate what is generally subsumed under the psychoanalytic terms, conscious and unconscious. I prefer the term awareness since it seems to reflect more closely the psychological quality of the phenomena whereas consciousness seems more appropriate for the designation of the state of physiological alertness of an organism.

---

[1] This paper was facilitated by Research Grant M-1206, National Institute of Mental Health, Public Health Service.

The most frequent definition of awareness and unawareness is in terms of verbal report. Awareness is equated with the ability to verbalize and unawareness with the inability or lack of verbalization. Freud (1925) in his extensive writings can be found on occasion to have defined awareness and unawareness in this manner. Such a distinction has been popular with other writers on the subject with Dollard and Miller (1950) most explicit. They define unawareness as the unverbalized or the unlabeled and their distinction between aware and unaware processes follows from their conception of the role of verbalization in the individual's adjustment. Verbal labels are assumed to increase the precision of discrimination and to mediate the transfer of training through the cue producing characteristics of the labels.

The popularity of a definition of awareness in terms of verbal report is understandable. It creates the illusion of operational precision and helps to mask the fact that many of the more important characteristics of awareness are avoided by the definition. Such an equation might seem a simple solution to the definitional problem but it is to be noted that the operational meaning of such a definition in any given experimental situation hinges critically upon a number of considerations. Among these we need to consider the adequacy of the questioning of the subject (*S*), the motivation of the *S* to respond with the care and precision that is required, the care taken to assure that the *S* understands what is being asked him, consideration of the effects of the interrogation itself upon the delicate process of awareness, and most importantly an adequate schema for classifying the *S*'s verbalizations along relevant dimensions. Assuming the last criterion can be met, a further requirement is some system of scaling or grading the *S*'s verbalizations along these dimensions in terms of their accuracy or specificity. It is not sufficient, as is too often the case, for the experimenter (*E*) to make a general statement that interrogation of his *Ss* revealed that they were unaware. It is the *E*'s decisions with respect to the above considerations that constitute the adequacy of his operational definition of awareness and any value the concept has depends upon the similarity of these operations from one experiment to another.

While attention to the above requirements constitutes a minimum standard for a definition of unawareness, it also is to be noted that there are some basic objections to the equation of awareness with a verbal report. There is nothing profound in noting that language and words are highly abstract symbols and bear little or no physical resemblance to the objects, events, and relationships that they denote. A definition of awareness in terms of verbalization places a heavy burden upon the adequacy of language to reflect the richness of perceptual experiences and images. The equation of awareness with verbal report leads to a similar kind of confounding of response and concept that has occurred with perception and the responses from which perception is inferred (Garner, Hake, & Eriksen, 1956).

There is also the question of how such a definition would handle voluntary responses other than verbal report. Thus in an experiment on complex perceptual discrimination an *S* may be at a loss to describe verbally the stimulus that has been presented to him but he can accurately choose the correct stimulus in a multiple-choice situation. Similarly, one's verbal description of the facial appearance of another individual may be wholly inadequate in enabling another to recognize this person among a crowd. However, if the person giving the description possessed a mediocre artistic talent, it would be possible for him to draw a sketch of the person described which would greatly increase the chances of the individual's being recognized. In these situations, if we adhered strictly to a definition of awareness in terms of verbalization, it would be necessary to say that the *S* was not aware of what the complex perceptual stimulus was like or of the described person's appearance. However, most of us will feel that this is failing to capture what we mean when we describe lack of awareness.

An equation of awareness with verbal report also would exclude from consideration a large body of experimental data that has supposedly dealt with the problem of awareness. As Adams (1957) has pointed out, one of the most readily demonstratable experimental arrangements supposedly showing unawareness is the following. In a psychophysical threshold determination if the *S* is instructed to give a confidence rating along with his judgments, it is frequently found that judgments which the *S* reports as pure guesses show above chance discrimination. If we equate awareness with verbal reportability, however, we would have to classify this as discrimination with awareness since the *S* is making a discriminated verbal report.

I have raised these objections and problems concerning the definition of awareness, not because I have a solution for them or a better definition to offer, but solely to insure that the issues raised in the remainder of this paper will be evaluated in terms of the limitations of our current definitions. I believe that in view of our present knowledge in this relatively unexplored research area, a definition of awareness in terms of a verbal report is a beginning point. But as was pointed out above any value such a definition has depends critically upon the operational specifications used in assessing the *S*s' verbalizations or lack thereof.

DISCRIMINATION WITHOUT AWARENESS

The question of discrimination without awareness is two separate questions with quite different implications. When we talk about discrimination without awareness we might mean: (*a*) is there a more sensitive (accurate) discrimination that occurs at levels not available to awareness; or (*b*) are we aware of all the discriminations that we make. The first question concerns the possibility that our nervous system differentially responds to stimuli or cues whose intensity is below that where a differentiated or discriminated verbal report can be obtained (subliminal discrimination). The second question concerns the possibility of differential responses to stimuli or cues above thereshold but which upon questioning the *S* is unable to state that he used these particular stimuli as cues for his responses (incidental discrimination). These two questions are seldom distinguished and it is not uncommon to find a psychologist arguing the affirmative answer to the first question by citing experimental data that only demonstrate the latter.

When discrimination at levels below that available to awareness is explicitly considered, it is often assumed that the discrimination occurring at these levels is only of a rather gross type. Cason and Katcher (1933) give a very plausible neurophysiological speculation in terms of how such subliminal stimuli could operate. They state:

> Before a conscious activity comes into existence, the energy that is produced or liberated in the body must be a certain required strength. The nervous energy frequently leads to some bodily effect even though its strength is not sufficient to produce conscious activity. It seems reasonable to believe that energy of this subliminal strength can be connected with some response by the learning process and if this could be demonstrated the theoretical implications would be important (p. 831).

Lazarus (1956) in defending his interpretation of the subception effect has suggested that discrimination at levels below that available to awareness as measured by the GSR, is capable of making gross discriminations between threatening and nonthreatening stimuli. Osgood (1957) has made a similar suggestion for interpreting perceptual sensitization and defense data. He conceives that an overall affective evaluation occurs early in the perceptual process and

> that hints provided by the incomplete information in the stimulus display, by eliciting the dominant evaluative (emotional) components of mediating reaction which feed cues back to the perceptual system, may produce either facilitation of covaluent integrations (vigilance) or disorganization of integrative mechanisms (defense).

He further suggests that this process may be mediated by lower brain centers.

Conceiving of discrimination at levels below awareness as only a rather gross type of discrimination has a certain plausibility to it. It avoids the concept of the "unconscious" as a sort of supernatural entity with mysterious powers. On cursory examination it would even appear to have experimental support. Binder (1955) among others, has pointed out that one or more of the characteristics of a stimulus may be discriminated in the absence of sufficient information of the remaining characteristics to permit specific identification. Thus, at a brief exposure we may be able to discriminate the color of a stimulus without perceiving its texture or its shape, and, if we can discriminate the color of a stimulus without also perceiving the texture and shape necessary for identification, why couldn't we be expected to discriminate the affective value of a stimulus in the absence of a specific identification?

The difficulty with such a conclusion is this. Generally, the *meaning* of a stimulus is conveyed not by one or two physical characteristics but by the total aggregrate of characteristics. In the case of words, which are frequently used as stimuli in perceptual defense studies, the affective meaning is conveyed only by the perception of a complete word, not by isolated characteristics such as word length or that there is a tall letter at the beginning of the word. To discriminate the word shot from a vulgar four letter Anglo-Saxon word would require a very precise discrimination as to a subtle form difference in only one of the four letters. For discrimination of the affective meaning of a threatening stimulus to occur, some place in the nervous system the complete word would have to be identified. So, we find that at what first seems to be a plausible assumption of only very gross discrimination at levels below awareness, is in fact an assumption of very precise discrimination, provided it is to be used to explain such phenomena as perceptual defense.

Experimental Studies of Subliminal Discrimination

What experimental evidence is there of a capacity for discrimination at levels below that available to verbal report? One can distinguish four basic methodological approaches to this problem. The first consists of attempting to condition some "nonvoluntary" response such as pupillary contractions or GSR to a stimulus presented with an intensity that has been previously determined to be subliminal. The second approach is to condition a GSR or similar nonvoluntary response to a stimulus at supraliminal levels and then use this response to assess the threshold for the stimulus. The third methodology, which is sometimes combined with the other two is to condition the GSR or some nonverbal response (or assume prior conditioning of this response) to a stimulus that is then presented around threshold values. The nonverbal response is then observed to determine whether it shows greater than chance relationship with the stimulus on those trials when the verbal judgment was incorrect. The fourth methodology is to present a stimulus, frequently a complex one, at a level assumed to be subliminal and then to look for the effect of the stimulation in subsequent behavior such as dreams, images, or free associations. I shall not attempt to summarize here all the different studies using these methodologies but instead select representative experiments which illustrate the general findings and the methodological problems encountered that limit the interpretation of the data.

Subliminal Conditioning

The evidence for subliminal conditioning is for the most part negative or at best controversial. One of the more intriguing studies is that of Baker (1938) who reported that the pupillary light reflex could be conditioned to a tone below the absolute threshold. Using the below threshold tone as a conditioned stimulus (CS) and change in illumination as the unconditioned stimulus (US), he was successful in establishing a conditioned response to the subthreshold tone within three trials. In addition, responses conditioned to the subthreshold sound generalized less than responses conditioned to suprathreshold values. They were also more stable and resistant to extinction.

In a repetition of the study, Metzner and Baker (1939) obtained essentially the same results, but other investigators have been unsuccessful in attempts to duplicate this experiment (Hilgard, Miller, & Ohlson, 1941; Wedell, Taylor, & Skolnick, 1940). Wedell, Taylor, and Skolnick found no evidence of conditioning in any of the *S*s they employed and a number of careful investigators have been unable to obtain pupillary conditioning even to suprathreshold stimulus values (Hilgard, Dutton, & Helmick, 1949; Steckle, 1936; Steckle & Renshaw, 1934; Stern, 1948).

In a recent experiment Taylor (1953) claims to have demonstrated subthreshold conditioning, but here again this interpretation of his results is open to serious question. He presented six different geometric figures tachistoscopically one at a time to determine a level of exposure where the *S*s were unable to identify at better than at chance level which specific geometric figure was being presented. He was successful in obtaining GSR conditioning at an exposure level where the *S*s were unable to correctly identify the different geometric figures.

While Taylor interprets his data in

terms of subthreshold discrimination, there is a more simple and plausible explanation of his findings. While his *S*s may have been unable to identify which geometric form was being exposed, this in itself is insufficient evidence for assuming that all six geometric forms at this level of exposure resulted in the same perceptual experience. It is more than likely that different perceptual experiences were associated with each of the six geometric forms at the level of exposure selected by Taylor. This is not to say that the information the *S* received was sufficient to specify what form was being shown, but merely that different experiences or perceptions were associated with the different forms. For example, the exposure of a square might have led to a description on the part of the *S* of a perception of a dark smudge that seemed to sort of expand in the field. A free report of his perception following exposure of the triangle might have led to some description such as "I had the impression of sort of a slanted line moving across the field." While these perceptions would not be sufficient to enable the *S* to label which figure was being presented, they nonetheless are distinguishable perceptions and therefore capable of differential conditioning or even differential verbal response provided proper feedback is given.

In Taylor's experiment a necessary control would have been to determine whether the *S*s reported the tachistoscopic exposure of a square and a triangle as identical perceptual experiences or as different. By asking the *S* to say "same" or "different" following successive exposures of geometric figures, one could have determined whether differential perceptual experiences existed at the verbally distinguishable level.

The possibility of subliminal conditioning of finger retraction was thoroughly and carefully explored by Newhall and Sears (1933). Using electric shock as the US and a light as the CS, these investigators were successful in obtaining conditioned finger retraction to light intensities at the absolute threshold (defined as the 50% point using the method of constant stimuli). Threshold determinations were also obtained throughout the experiment. Some of their *S*s were found to give conditioned finger retractions to light intensities below the absolute threshold (four of their seven *S*s gave a total of five conditioned finger retractions in the sigma intervals of − .5 to − 1.0).

Newhall and Sears were properly cautious in interpreting their data in terms of subliminal conditioning. They pointed out that while some conditioned finger retractions were obtained at values below the absolute threshold, the threshold in their study was defined as the 50% point. Thus by definition of the threshold, the *S*s verbally would report on a number of occasions having seen the light at values that actually fall below the arbitrarily defined threshold.

Further confirmation that there was no super discrimination by unaware process in their experiment is found when the conditioned response is itself used to estimate the absolute threshold for light. By using the latter sessions in the experiment when conditioning had been well established in most *S*s, it was possible to estimate the thresholds from the frequency of conditioned finger retractions. It is interesting that the absolute threshold computed from the frequency of conditioned responses corresponds almost exactly to the value obtained by using the verbal judgments.

The most recent and perhaps the most thorough investigation of subliminal conditioning appears to be that of Wilcott (1953). He attempted to condition GSR to subthreshold auditory intensities at several different frequency levels. His *S*s were given preliminary training on threshold determination and the stimulus intensity selected for subthreshold conditioning was the *highest intensity* that the *S failed* to report hearing during the threshold determination sessions. While he was successful in GSR conditioning to suprathreshold intensities, no evidence was found for conditioned GSRs in response to the subthreshold intensities at any frequency level. Interestingly enough, Wilcott's experiment appears to be the only one using a subliminal CS that was at a low enough intensity so as not to yield better than chance verbal reports.

From this brief survey of the literature we can see that the evidence for discrimination at levels below those available to awareness is not very impressive. In the cases where studies have obtained positive results, other investigators have been unable to duplicate the findings. In the cases where slight evidence for subthreshold conditioning has been obtained the results seem to be more readily accountable by other explanations.

There are several sources of artifact that need to be carefully considered before interpreting the data of an experiment in terms of conditioning or discrimination at levels below those available to awareness. Of major importance is the definition of a threshold. It must be borne in mind that our thresholds are arbitrary statistical definitions. In the typical psychophysical study an *S* is asked to report "yes" or "no" concerning the presence or absence of the stimulus. Commonly the 50% accuracy level has been accepted as a threshold. It is to be noted, however, that this procedure insures that the *S* is discriminating better than chance at a stimulus intensity below the arbitrary threshold designation. If a blank stimulus interval is included in the threshold determination, the *S*'s *false alarm rate* or his tendency to say "yes" in the absence of any stimulation can be observed. In a good *S* this false alarm rate may be 10% or less. Thus, any deviation from this 10% point in terms of positive answers or reports that he perceives the stimulus indicates he is now beginning to perceive at better than a chance level. At a stimulus intensity that the *S* reports perceiving 25% of the time, we are below his arbitrarily defined threshold value but at the same time we are 15% above his chance performance level. To demonstrate that a conditioned response can be obtained 15% of the time at this stimulus intensity level is hardly evidence for concluding discrimination below verbal awareness.

Further sources of error in the interpretation of subliminal conditioning studies revolve around the method used in obtaining thresholds and the instructions and training given the *S*s. We shall not go into these problems in detail here since the interested reader can consult the excellent discussion of these problems by Goldiamond (1958). However, we may note that the "threshold" obtained for an *S* depends upon a number of variables. Most important are: (*a*) the instructions and expectations of the *S*s; (*b*) the penalties for false alarms and for failure to detect a real stimulus; (*c*) the training given the *S*; (*d*) the use or lack of use of a ready signal; and (*e*) whether or not a forced-choice method is employed (Blackwell, 1953). Most typically experiments on subliminal stimulation have given only cursory attention to the operations by which they have concluded that a given stimulus was "subliminal." Few, if any, would meet the exacting requirements of modern psychophysics.

Assessing Thresholds by Conditioned Responses

A second method for investigating the question of a more sensitive discrimination that occurs at levels below awareness is to condition some response other than verbal to a stimulus and then observe whether this response generalizes to values of the stimulus at threshold intensities. One could condition a GSR to a suprathreshold stimulus and then observe the frequency of conditioned responses as the stimulus was varied in intensity around the threshold level. The frequency of conditioned responses as a function of the intensity of the stimulus could be used to compute a threshold that could then be compared with the threshold obtained by verbal report. The experiment by Newhall and Sears described above provided an example of this method of threshold determination.

One of the more interesting experiments of this type is reported by Davis (1950). Recordings were made of the action potentials in several different muscle groups in response to a sound presented above and below threshold. Increases in muscle tension occurred to stimulus intensities that the *S*s did not report hearing. Unfortunately, these data can be taken as only suggestive since the method of threshold determination followed by Davis was not very rigorous. The *S*s were not given preliminary training nor were they alerted for the intervals in which the stimulus might appear. In addition the instructions given were not of the type that generally yield a most sensitive verbal threshold.

Even if replication should substantiate Davis' finding that muscle tension varies at levels of stimulation below those capable of differential verbal report, there is still the question of what meaning this has for behavior. For responses by some part of the body to external stimulation to be important psychologically it needs to be shown that these responses are capable of learning and serving as cues for behaviorally more important responses.

Recently Dulany and Eriksen (1959) reported an experiment in which the GSR threshold for brightness was systematically compared with verbal response thresholds determined by two methods—a modification of Blackwell's forced-choice as well as the more typical presence or absence report from the *S*. *S*s were trained for 10 days prior to a conditioning session. The GSR was then conditioned to a suprathreshold intensity of the test light with electroshock to the ankle as the US. Following conditioning, 3 more daily sessions were employed to establish the GSR threshold as well as a concurrent forced-choice discrimination threshold. To make the thresholds obtainable from the GSR and the verbal response directly comparable, the test patch could occur in either of two intervals designated by an auditory signal. *S*s were required to report verbally which interval the light occurred in and their GSRs were concurrently recorded. If the GSR to the first interval was greater than the GSR to the second, this was considered the GSR's choice on that particular trial.

While these investigators were successful in obtaining stable GSR thresholds in their *S*s, there was no indication that the GSR gave a lower threshold than did either of the two verbal report methods. In fact, the verbal reports in most *S*s were found to show a deviation from the chance base line at lower stimulus intensity than the GSR's. Also, it was noted that while the verbal reports obtained an accuracy level of 100% at higher stimulus intensities, the GSR never achieved this high level of discrimination. This latter observation is important in that any attempt to claim greater sensitivity for the GSR would have to assume that somehow a response which is less sensitive or discriminating at above threshold intensities metamorphoses over the stimulus range to become more sensitive at low stimulus intensities.

It is to be noted that the Dulany and Eriksen experiment does not prove that the GSR is no more sensitive than the verbal report in obtaining a discrimination threshold. It is quite possible that the many sources of error that occur in GSR recording whether due to the equipment or to the nature of the response itself may prevent a fair comparison of the two response systems. Also, there is a possibility that *S*s have had greater practice in using the verbal responses as opposed to the GSR for such a discrimination. These, at the present time, appear to be unanswerable possibilities but it is to be noted that the care taken in training GSR and verbal response thresholds in this experiment was much more extensive than that employed in experiments that have used GSRs as an indicator of unconscious perception.

The Subception Effect

There is a third class of experiments that is erroneously but nonetheless often considered to show discrimination at levels below that available to awareness. These experiments make use of the same experimental arrangement as that employed in the previous approaches. A nonvoluntary nonverbal response such as a GSR is conditioned to some suprathreshold stimulus or stimuli and then the stimuli are presented at about threshold value. The *S*s are required to give a verbal response and the incidence of the concurrent conditioned response is noted. However, in these experiments the conditioned response is not used to compute a threshold. Instead the data are analyzed in terms of whether or not a conditioned response occurs more often than chance would indicate in the absence of a correct verbal response. Lazarus and McCleary (1951) in their subception experiment were among the first to use this type of analysis as evidence for discrimination without awareness.

They conditioned GSRs to 5 of 10 nonsense syllables and then exposed the nonsense syllables individually in a tachistoscope at a duration level where verbal recognition was about 50% correct. The average GSR to shock syllables was found to be greater than to nonshock syllables on those trials where the *S*s were unable to give the correct verbal identification of the syllables. They interpreted their data in terms of subception: "a process by which some kind of discrimination is made when the *S* is unable to make a correct conscious discrimination."

While nearly all of the discrimination without awareness experiments that we have analyzed so far would have lent their data to this type of comparison, the novelty of the Lazarus and McCleary approach was the introduction of this method of analysis. It is not the same as determining a differential threshold for the GSR and the verbal response. Lazarus and McCleary did not compute separate thresholds. While they report that the tachistoscopic level was set so as to obtain approximately 50% correct verbal recongnitions, they do not report the differential accuracy of the GSR on the differential accuracy of the GSR on the differential accuracy of the GSR on classifying shock and nonshock nonsense syllables. Such information is important to our present discussion and can be estimated from an unpublished replication of their study conducted in our laboratory. This replication was successful in duplicating the subception effect. But a comparison of the relative accuracy of the GSR and the verbal report shows that if an observer wished to predict whether a shock or a nonshock nonsense syllable had been presented to the *S* and the observer could use only one of the *S*'s responses, the verbal report or the GSR, he would have been appreciably more accurate using the *S*'s verbal report.

At the exposure level used in the replication approximately 45% of the verbal reports correctly identified the specific nonsense syllable. The most optimum weighting of GSR values gave approximately 60% correct classifications into shock or nonshock categories, but the chance expectancy for correct nonsense syllable identification for the verbal response is 10%, whereas for the dichotomous classification of shock vs. nonshock by the GSR, chance level is 50%. An observer attempting to determine whether the *S* had seen a shock or a nonshock syllable would have been correct in his predictions 60% of the time using the GSR but using the verbal response he would have been correct 45% of the time plus one-half of 55% for a total of 72.5%. (This assumes that the *S* would randomly distribute his errors on the verbal response between shock and nonshock syllables.)

While the Lazarus and McCleary experiment does not demonstrate that the threshold of a GSR response is lower than the threshold for verbal identification, nonetheless, their finding that in the presence of an error in the verbal response system the GSR response has better than chance accuracy has some important implications for a concept of behavior without awareness. Elsewhere the present author (1956a) has analyzed the subception experiment in terms of a multiple concurrent response model. In the Lazarus & McCleary experiment the *S* gave two concurrent responses to a stimulus—a verbal response and a GSR. The obtained subception effect consists of a partial correlation between the stimulus and GSR when the verbal response is held constant. Conceived in this way it is possible to specify the general circumstances under which such a partial correlation would be obtained.

The necessary and sufficient conditions are fulfilled when both responses are independently correlated with the stimulus but less than perfectly, and less than perfectly correlated with each other. The independent correlation of the two responses with the stimulus means that both the verbal response and the GSR are responses to the stimulus as perceived rather than the GSR being a response to the verbal classification of the stimulus. Less than a perfect correlation between the verbal response and the GSR can be alternatively stated as saying that these two response systems have noncorrelated error terms. Thus, variables which may affect the accuracy of use of verbal responses may not concurrently affect the accuracy of the GSR and vice versa.

Formulation of the results of the subception experiment in terms of partial correlation enables us to state specifically just what the subception experiment demonstrated. It succeeded in showing that both the verbal response and the GSR are each, at least to a degree, independently correlated with the stimulus and are partially independent of each other. In other words, the GSR is, at least partly, determined by the stimulus and the intervening perceptual process and not solely by the verbal response. This relation can best be seen in terms of

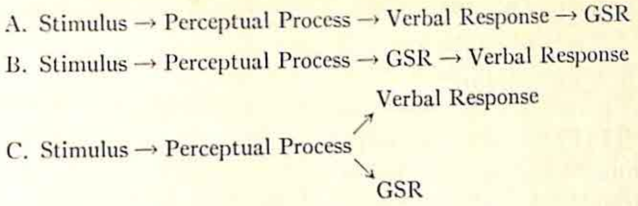

A. Stimulus → Perceptual Process → Verbal Response → GSR
B. Stimulus → Perceptual Process → GSR → Verbal Response
Verbal Response
C. Stimulus → Perceptual Process
GSR

FIG. 1.

the diagrams in Figure 1. If the relation was as depicted in A, then it would have been impossible to have obtained a subception effect. A partial correlation of the GSR with the stimulus holding constant the verbal response could not have been obtained. Similarly if the relationship was as depicted in B, it would be impossible to find a significant correlation between the verbal response and the stimulus when the GSR response was held constant. While the Lazarus and McCleary data are not presented so as to show whether or not a significant partial correlation between the verbal response and the stimulus existed when the GSR was held constant, in the replication of their experiment mentioned above a significant and appreciable correlation of this nature was obtained. This would seem to rule out the relationship as diagramed in B. This leaves only the relationship as diagramed in C where we have depicted the GSR and the verbal response as at least partially parallel response systems both independently responding to the perceptual process.

The author has carried out a series of experiments in his laboratory to test this partial correlation formulation and to see what implications it has for a concept of behavior without awareness. The subception effect has been successfully obtained in studying the concurrent GSR and verbal response generalization to a series of different size squares (Eriksen, 1956b). Here it was found that when the verbal response was held constant, a generalization gradient for the GSR as a function of square size was still obtained. In succeeding experiments verbal responses together with a concurrent lever movement response were studied (Eriksen, 1957, 1959). Again, if the instances where the verbal judgment or response was incorrect were analyzed the concurrent lever movement response showed better than chance accuracy. The effect also has been successfully obtained where the concurrent responses were an affective judgment of pleasant and unpleasant together with a specific verbal identification of tachistoscopically exposed stimuli (Eriksen, Azuma, & Hicks, 1959). In all of these studies a significant partial correlation was obtained between either response and the stimuli with the other response held constant. In no case, however, was it found that the nonverbal judgment gave a lower threshold or more accurate discrimination than did the verbal judgment.

These studies were also successful in indicating several variables that constitute the sources of noncorrelated errors between two concurrent response systems. Thus, it was found that distracting signals will introduce increased error into GSR responses without affecting the accuracy of concurrent verbal responses and similarly the number of competing alternative verbal responses affects the error of this judgment without concurrently influencing the accuracy of the GSR (Eriksen, 1956b).

On the basis of the above evidence the subception effect as an effect is real enough but it has nothing to say and offers no proof concerning the capacity of the GSR to reflect a more sensitive discrimination than that which is available through the *S*s' verbalizations. Instead the subception effect focuses attention upon the limitations of verbal responses in conveying the individual's perceptual experiences. By formulating the effect in terms of partial correlation in a concurrent response system model, we have the advantage of being able to direct our attention upon sources of noncorrelated errors between two or more concurrent response systems. We have already indicated some of the factors that will lead to increase in noncorrelated errors between concurrent responses and it is interesting to speculate on the effect that variables such as differential reward and punishment might have upon decreasing the correlation between verbal and certain nonverbal responses.

To say that the subception effect does not provide evidence for discrimination at levels below awareness is not to deny that the phenomenon is important to our general concept of awareness. There is reason to suspect that a considerable amount of the evidence that the clinician finds for unconscious processes is in the nature of discrepancy between concurrent responses. The clinician may note the *S*s' autonomic responses and evidences of emotion do not correspond, i.e., do not correlate very well, with his verbal statements. This lack of correlation is a basis for the clinician's inference of unconscious processes or defensive mechanisms at work. A concurrent response system model would seem the efficient way of conceptualizing this situation and suggesting ways of experimentally investigating the clinical phenomena.

Indirect Effects of Subliminal Stimulation upon Behavior

A recent experiment by Dixon (1958) has introduced an interesting and novel method for demonstrating effects of stimulation below the awareness threshold. Dixon determined the brightness threshold for a vertical and a horizontal line by the ascending method of limits. Having obtained a threshold Dixon told his *S*s that he was going to expose words to them at a brightness level where they would not recognize the words. He then exposed 10 words or phrases (6 were emotional in connotation and 4 neutral) at a brightness level .3 log ft. lamberts below the threshold obtained for the vertical and horizontal lines. The *S*s were told to give a word guess to each exposure even though they would not see the word. Each of the 10 words or phrases was exposed four times during a session and then at a later session the *S*s were presented the list of stimulus words and their responses were read to them. They were instructed to try to match their responses with the stimulus words using any type of associative connection they wished. Dixon's hypothesis was that the words exposed below threshold would have been unconsciously perceived in terms of their meaning and would thus influence the associations or responses the *S* made. Five of the seven *S*s gave results in this predicted direction. In fact, when the individual probabilities for each *S* were summed through *S*s the results were significant at a high level.

Dixon's approach to the study of discrimination without awareness is an interesting departure in methodology but subsequent work indicates his results are artifactual. Fuhrer and Eriksen (1960) in an extensive replication of the experiment confirmed that *S*s could match associations given to stimuli that had been exposed at illumination levels too weak for correct identification. However, systematic employment of controls showed this effect to be a result of structural cues such as word or phrase length rather than "unconscious perception of stimulus meaning."

Fuhrer and Eriksen were unable to obtain significant or suggestive results at brightness levels of stimulus exposure below the threshold for vertical and horizontal lines obtained by the ascending method of limits or at the threshold obtained by a modification of Blackwell's forced-choice technique. However, when exposure was at a brightness level where *S*s could detect the presence of a stimulus but not identify it, significantly better than chance matching occurred. That the matching by these *S*s was due to structural cues rather than the perception of meaning was shown by the data from a control group. Using the same exposure level but now exposing the stimulus words and phrases upside down and backwards, the control *S*s were also found to give significantly better than chance matching of their associations to stimulus words. There was no significant difference in performance between this control group and the previous experimental group which had used the same exposure level. Since it seems extremely implausible that unconscious discrimination is capable of perceiving the meaning of words that are flashed upside down and backwards at very low illumination levels, it would appear that the performance of both groups is attributable to the cues perceived concerning the length of words or phrases.

But to check upon this possibility a further control was employed. For the experimental group a control list of words similar in structure but different in meaning to the actual stimulus words was constructed. Thus, for the stimulus phrase "male sex organs," the structurally similar control phrase was "made six organs." In the second session, when the experimental *S*s were asked to match their associations to the stimuli, they were also presented with this control list of structurally similar words together with the experimental words. Analysis of the results revealed no tendency for this group to match their associations with the correct words over the structurally similar control words.

The experimental work so far offers no convincing evidence that the human observer can respond differentially to stimulation that is of an intensity level below that at which a discriminated verbal report can be obtained. In other words, when adequate controls are provided, no response has been found to yield better absolute discrimination than a verbal response. However, there are most likely circumstances where a nonverbal response may be a better indicator than verbalization. No psychophysicist, for example, would expect an experimental arrangement where the *S* pressed a button directly in front of him to indicate "No" but had to get up and walk across the room to press a button to indicate "Yes" to yield as low a threshold as a situation in which both buttons were directly in front of the *S*.

Such response biasing can occur in less obvious ways. It is quite possible that anxiety conditioned to certain responses may bias their occurrence. Similarly, artificial restrictions on the number of verbal labels available can have a similar effect as well as deliberate response suppression due to social taboos associated with the response. But it is to be noted that under these circumstances one would not conclude evidence for a super unconscious discrimination of stimulus intensities too weak to elicit verbal discrimination.

AWARENESS AND THE SUBJECT'S SUBJECTIVE CONFIDENCE

There is one other experimental approach to the problem of awareness that deserves brief consideration. It differs from the other methods in that a discriminated verbal report is not used as the criterion or definition of awareness. Instead awareness is defined in terms of the *S*'s statement as to whether or not he feels he has perceived the stimulus or whether he feels he is merely guessing. The evidence consists of showing that when the *S* expresses zero confidence, e.g., subjectively pure guessing, he nonetheless responds correctly to the presence or absence of the stimulus at better than chance level.

Peirce and Jastrow (1885) were among the first to report this phenomenon. Working with the difference limen on weight judgments, Jastrow's data indicate that in the presence of a subjective report of "no confidence," his judgments of heavier or lighter were still correct significantly better than chance. Since then this phenomenon has been a frequent finding in a number of experiments. In fact, as Adams (1957) points out, it is a stable enough phenomenon to permit its use as a classroom exercise. Representative of experiments obtaining results on this aspect of awareness are those of Fullerton and Cattell (1892), Miller (1939), Coover (1917), and Williams (1938).

Although these experiments are frequently and uncritically used to support contentions concerning discrimination without awareness where awareness is defined in terms of a discriminated verbal report, it is to be noted that these experiments do not fit this definition. Here a discriminated verbal report is obtained. The evidence for unawareness is in terms of the *S*'s report of lack of confidence in his verbal judgments. While this subjective confidence in his reports is an interesting problem, it has no operational similarity to the type of experiments we have considered above. An equation of awareness with verbalization, by definition, means that any verbal judgment is one made with awareness irrespective of whether the *S* has little or no confidence in the judgment.

The major difficulty with experiments of this kind is that they place upon the individual *S* the responsibility for establishing the criterion of awareness. We have no way of objectively knowing when *S* A says a given judgment was a pure guess and *S* B makes a judgment and states it was a pure guess that the two *S*s are talking about the same phenomena or that they are using the same criteria for guessing.

In work in our laboratory we have found very marked individual differences for the intensity range between where an *S* shows better than chance guessing in his responses and the point where he begins to report confidence in his judgments. For some *S*s the stimulus intensity where responses are better than chance almost coincides with their confidence threshold. For other *S*s there is a great range between where guesses are above the chance level and the point where confidence is reported in these judgments. While research on this particular aspect of awareness may be indeed fruitful, nonetheless, at the present stage of our knowledge we are not justified in lumping together data from these experiments with those that have been discussed in the preceding sections. Not only are they operationally dissimilar but more importantly, these marked individual differences indicate that as long as we rely upon the *S* himself to report whether or not a verbal judgment was made with awareness, we are in fact using as many different criteria of awareness as we have experimental *S*s.

BEHAVIOR WITHOUT AWARENESS TO ABOVE THRESHOLD CUES

While we have been unable to find evidence for a supersensitive discriminating unconscious, the evidence that behavior can be directed by above threshold cues of which the *S* is unaware, is not only more plausible but somewhat more substantial. Common sense tells us that we are constantly utilizing cues of which we are unaware in our perception of depth, of shape and size constancy, as well as a number of cues that guide such complex motor habits as piano playing and driving automobiles. One of the factors making the elimination of such habits as fingernail biting or other mannerisms hard to eliminate is that they tend to occur without our awareness. Experiments appear to have demonstrated not only the use of cues without awareness but also the lack of awareness in some instances for the response the *S* is making, for the reinforcements controlling the response and in some cases the contingent relationship between cues, responses, and reinforcements.

Many of the experiments providing evidence on the effects of cues without awareness were considered by the *E*s as demonstrating effects of subliminal stimulation upon behavior. As will be seen the use of subliminal in connection with these cues was gratuitous since the particular experimental methodology did not contain the necessary controls or operations to permit the classification of these cues as subliminal. In fact, on the basis of other evidence there is every reason to suspect that these cues, when effective, were above a chance discrimination level for verbal report when assessed by proper psychophysical techniques.

The methodology employed in experiments in this area makes use of one or another of a few basic techniques. In most cases the cue is presented without forwarning the *S* that it will be present. Typically it is presented at an intensity so low as not to command attention. In other experimental arrangements the cue may be of a more obvious nature but the *S* is prevented from attending it by misdirection as to the purpose of the experiment or misdirection concerning the nature of the experimental task. In fact, misdirection of attention seems to be a crucial consideration in experiments of this nature, suggesting that perhaps we are dealing more with a concept of attention than one of awareness, if indeed the two concepts can be distinguished.

One may distinguish two separate lines of investigation using incidental cues. The first is concerned with the effect of incidental cues upon behavior that was learned outside of the experimental situation. For example, the effect of the incidental stimulation of the word "angry" upon the interpretation of ambiguous profiles might be studied.[2] The other approach is concerned with the problem of learning without awareness. The first approach we will term performance without awareness to distinguish it from the problem of learning without awareness.

Besides sharing the common use of the technique of incidental stimulation both lines of research investigation also share a common problem in the definition of awareness. Both must rely upon the interrogation of the *S* concerning his awareness in the experimental situation. As was pointed out earlier in this paper such interrogation is fraught with difficulties. The value of the concept of awareness depends upon the adequacy with which the *E* surmounts these problems. The operational meaning of awareness will vary markedly from one experiment to the next depending upon the adequacy of the *E*'s assessment of the *S*s' verbalizations and the precautions taken to insure that they are valid and meaningful.

---

[2] I am indebted to George S. Klein for the term "incidental stimulation."

Performance without Awareness

There have been several experiments demonstrating that unaware cues are capable of producing perceptual illusions. Of the series of studies utilizing the Mueller-Lyer Illusion (Dunlap, 1900; Titchner & Pyle, 1907), the one by Bressler (1931) appears to be the most thorough and best controlled. He found evidence of illusory effects in the judgment of the lengths of the two lines when the directional arrows were so faint that the *Ss* reported that they were unaware of their presence.

Miller (1939) and Perky (1910) projected geometric forms on the back of a ground glass screen at low intensities and found that the patterns presented were influential in determining the *S*s' choice of geometric pattern. The effect was found even when the *S*s were unaware that a pattern was actually being projected on the screen.

More recently investigators have made use of the metacontrast (Alpern, 1952) phenomenon as a method of presenting cues in such a way that the *S* would be unaware of them. Smith and Henriksson (1955) presented a stimulus consisting of fan shape lines, very briefly in a tachistoscope, and then, almost immediately after, flashed a square stimulus. While their *S*s did not report seeing the fan shaped lines, they did tend to see the square as trapezoid. Kolers (1957) has presented tachistoscopically sets of designs that required abstraction or conceptualization and found that if cues to the solution of the problems were flashed first, these cues were capable of influencing the type and direction of the solution even though *S*s reported they were unaware that the cues had been presented.

Using a modification of this metacontrast technique, Smith, Spence, and Klein (1959) presented a line drawing of an expressionless human face to their *S*s. The face was continuously presented to the *S*s except for brief interruptions where the word happy or angry was briefly flashed. They report suggestive evidence that the *S*s' interpretations of the face were more pleasant following exposures of the word happy than following exposures of the word angry (12 of 20 *S*s mentioned more pleasant characteristics of the face following exposures of the word happy than they did following exposures of the word angry). Exposures of the word angry however, had little or no effect upon the *S*s' interpretations of the face.

It is to be noted that the experiments described above in this section have been interpreted as demonstrating the effects of *subliminal* cues upon perception. If by subliminal we mean that the intensity of the cues was too weak to elicit a verbal report threshold under careful psychophysical conditions, then none of the above experiments permit interpretation in terms of subliminal stimulation. In Bressler's experiment even the faintest lines he employed, when studied independently of the illusion, were found to yield better than chance discrimination by his *S*s. Vinacke's (1942) data clearly show that the Miller and Perky results are not obtainable when the geometric forms are presented at an intensity level below a rather crudely determined psychophysical threshold. Similarly the metacontrast effect seems to depend upon the first stimulus being at a level that would have been detectable were it not "washed out" by the second stimulation. And in the Smith et al. study, observations upon a control group of *S*s who had been alerted to the fact that words were being flashed indicated that most of the word exposures for the experimental group were above the level for detectability of the word. No evidence is presented to show that the presentation of the words happy and angry had a significant effect upon the interpretation of the profile when presented at durations below the detectability level.

Even when the above studies are considered as demonstrations of incidental rather than subliminal stimulation the conclusions must be accepted with caution. In most cases, the effects are slight even though statistically reliable. Careful replication of these studies is needed before confidence can be placed in the effect. And then, assuming replication substantiates the results, careful parametric analyses should be undertaken to explore not only the parameters of the experimental situation that create conditions where the *S* is inattentive to or unaware of the incidental stimulus but also studies oriented to studying the complexity of cues that can be utilized in this manner. Also, the effects of this incidental stimulation should be compared with the effects when the stimulation is actually detected in awareness.

There is another group of studies that would seem to fall under the classification of incidental stimulation but because of their particular nature deserve special consideration. These are the experiments based upon the Poetzl phenomena. In 1917 Poetzl reported a series of studies on his patients in which he exposed complex pictures tachistoscopically and instructed his patients to dream about the picture. He found that items in the picture which were not reported in intentional recall tended to occur in the dreams. Interest in this phenomenon was revived briefly in the early 1930's (Malamud & Linder, 1931) and then lay dormant until recently revived by Fisher (1954). An experimental report by Shevrin and Luborsky (1958) illustrates the methodology of research on this phenomenon as well as the general findings and methodological problems encountered. These investigators tachistoscopically presented a complex scene to their *S*s following which they asked for a description and drawing of what the *S* had seen. *S*s were then asked to bring in their dreams in a subsequent session. In this subsequent session further intentional recalls were obtained and the objects obtained in intentional recall were compared with the objects that appeared in the *S*s' dreams. Since many of the *S*s did not report dreams, they were asked to conjure up a mental picture or image which they were then asked to draw and to describe.

Shevrin and Luborsky report that a significant number of items and characteristics of the tachistoscopically exposed scene that were not mentioned in the intentional recalls occurred in the dreams or the *S*'s images. These findings were interpreted in terms of the psychoanalytical concept of day residues.

The complexity of these studies and experimental arrangements poses many methodological problems and gives rise to the possibility of numerous artifacts. One of the first problems encountered is the frequency or base rate with which different objects or items spontaneously occur in dreams and images independent of a prior pictorial stimulation. This is a particularly acute problem since these investigators used a rather loose criterion for determining whether objects that had appeared in the picture occurred in the dreams or images. Thus, if the picture contained trucks and the *S* in his dream or image mentioned a bicycle, this was scored as an additional memory recovery since both represented a means of conveyance.

A second problem involved in the interpretation of Poetzl phenomenon data concerns the previously mentioned problem of the *S*s' subjective confidence level. When instructed to recall or describe what had been seen in the tachistoscopically presented picture we may assume that the *S* has a certain veridical orientation. That is, he wishes to be correct. As we have previously seen, *S*s have different criteria for reporting what they believe to have existed. Some *S*s may not mention an item unless they have fairly high confidence that it was present while others may use a lower confidence level. However, when freed of the veridical orientation required by intentional recall and asked to image, most *S*s would be expected to have freer criteria for what they would report. Thus, items which they were not certain were in the picture or for which they had low confidence are more apt to appear in the image than under instructions for intentional recall.

A third problem is what we might call the contingent probability of items going together in our enviornment. To illustrate what I mean by this, let us consider an *S* who has no recall of having seen a boat or a wharf in a picture but he does describe a body of water in his deliberate recall. When he is now asked to image, it seems likely that his associations are apt to be conditioned by the thought of water. Thinking of water it is likely that associations might run toward boats and wharfs and other objects that have in his past experience tended to be associated with a body of water. Since these objects would also have a fair probability of having actually been in the picture, this would lead to an increase in the recalled items in the image condition which was spurious due to the past learned associations among objects in our environment. In subsequent work on this problem a satisfactory control for this possibility might be obtained by having control *S*s who do imaging after observing the experimental *S*s' drawings and descriptions of the picture. Another control possibility would involve the use of stimulus pictures composed of heterogeneous objects whose inclusion was determined by random numbers.

Learning without Awareness

On the basis of the large number of studies during the past few years on learning without awareness, it would seem that the phenomenon itself should be well established by now. No attempt will be made here to survey the large literature. This has been well done by others (Adams, 1957; Krasner, 1958). Rather, this research area is mentioned in order to show that the phenomenon is not as well established as is generally supposed and also to indicate some of the directions that future research in this area must take.

As was also the case with research on performance without awareness, learning without awareness studies have been plagued with the problem of adequate operational criteria for awareness. Many of the experiments in this area have been concerned primarily with operant conditioning and little or no concern has been given to the definition of awareness. In many cases the only criterion is the statement by the *E* that the *S*s were unaware or that the *S*s were unable to label the correct response class. A few studies have focused primarily upon the phenomenon of awareness and here the evidence for learning without awareness has not been impressive (Krieckhaus & Eriksen, 1960). Adams (1957) in his survey has pointed out many of the inadequacies in this field of research such as the lack of replication of experiments, and more particularly the failure of *E*s to adequately rule out the possibility of correlated hypotheses.

More recently Dulany (in press) has called into question many of the verbal conditioning studies by showing that the results could be attributable to correlated hypotheses and the transfer of previously acquired verbal habits. Dulany used what has become a typical verbal conditioning procedure, studying the influence of the *E*'s "mmm hmmm" upon the *S*'s plural noun output. He found a significant increase in plural noun output by his *S*s although none of them were able to verbalize the correct response category. In this sense, then, they were unaware. But Dulany analyzed the data from his *S*s further. He found that the only *S*s who showed a learning effect were those who were able to verbalize to the effect that the *E*'s "mmm hmmm" indicated to them that they were associating in the proper category. He further demonstrated that associations to plural nouns tend to be other plural nouns whereas associations to singulars do not show this tendency. Since the *S*s were interpreting the "mmm hmmm" as an indication they were to continue associating in that category and the "mmm hmmm" followed a plural, this effect would be sufficient to account for the learning shown by his *S*s.

Research on learning without awareness as well as unconscious phenomena in general has been characterized by uncritical acceptance. Diven's (1937) experiment on unconscious conditioning of anxiety has been extensively cited and used as the basis for considerable theorizing concerning psychodynamics. Yet as Lacey and Smith (1954) have so clearly pointed out the Diven experiment did not contain the basic data comparisons necessary for a logical conclusion of conditioning without awareness. Using a more sophisticated design, Lacey and Smith, however, reported that they obtained evidence of unconscious conditioning of anxiety. But here again, Chatterjee and Eriksen (1960) have shown that Lacey and Smith's evidence on conditioning may well have been due to an artifact arising from the method of computing conditioning scores. These latter investigators were in general successful in replicating the Lacey and Smith findings but due to a more careful assessment of the *S*s' verbalizable awareness, they were able to show that the autonomic conditioning was no more precise nor specific than the *S*s' verbalizations.

Perhaps our tendency to uncritically accept experiments on unconscious phenomena may be due to our firm belief in the existence of the unconscious. But it is to be noted that while we might firmly believe in the phenomena of classical conditioning this does not discharge an *E* from the duty of showing that his particular conditioning results are not attributable to the artifact of "pseudo conditioning." There would seem to be little doubt that a considerable amount of human behavior occurs without awareness of the behavior at the time of its occurrence but it is to be noted that this does not logically require that behavior is learned without awareness. It may well be that the awareness process is biologically adapted to be sensitive to features in the environment that are important to the organism and that any circumstances important enough to the organism to produce learning would not go undetected by awareness. This is not the same as saying that awareness is necessary for learning. Rather, it suggests that any situation where the cues and the reinforcements are salient enough to produce learning, will not escape detection by awareness.

It would seem that on the basis of the available evidence learning without awareness in human *S*s is not adequately proven. A considerable research effort is needed not only to arrive at acceptable operational criteria of awareness but to thoroughly explore the different types of learning situations to determine what if any kind of learning can take place without awareness. It may be that only simple conditioned responses are capable of learning without awareness or it may be that fairly high level concepts can be acquired. More attention in research needs to be paid to the awareness process itself, particularly in terms of criteria and independent correlates of awareness. Also, systematic parametric analyses of the experimental situation and arrangements where learning without awareness occurs must be undertaken. These latter studies should go far in suggesting an improved concept of awareness or attention.

SUMMARY AND CONCLUSIONS

This paper has surveyed the experimental work on discrimination, performance, and learning without awareness. The methodology employed in these experiments has been critically examined and the nature of the evidence evaluated. Some of the problems inherent in the definition of awareness have been considered in relation to the experimental work and an attempt has been made to show that there are a number of separate questions that have traditionally been lumped together under "unconscious" phenomena. Artifacts involved in experiments on subliminal perception and subception were pointed out and alternative interpretations for these experiments were suggested.

It was concluded that at present there is no convincing evidence that the human organism can discriminate or differentially respond to external stimuli that are at an intensity level too low to elicit a discriminated verbal report. In other words, a verbal report is as sensitive an indicator of perception as any other response that has been studied.

Consideration of the clinical evidence for the concept of the "unconscious" suggests that an assumption of "subliminal discrimination" is not required to account for the observed phenomena upon which the concept is based. Instead, the evidence noted by the clinician or student of personality would seem to be explainable in terms of either the demonstrable partial independence of different response systems to fragmentary perceptual cues and memory traces and/or to lack of awareness to above threshold cues.

A science of personality is not furthered by the frequent tendency of psychologists to discuss the "unconscious" with all the ambiguity and reverence that religions accord to the soul. There is great need to spell out explicitly the assumed characteristics of the unconscious and to search for explanations of so called unconscious phenomena in terms of more commonplace psychological variables. To do so may destroy the titillating mystery that the unconscious seems to hold but then that is the business of science.

REFERENCES

Adams, J. K. Laboratory studies of behavior without awareness. *Psychol. Bull.*, 1957, **54**, 383–405.

Alpern, M. Metacontrast: Historical introduction. *Amer. J. Optom.*, 1952, **29**, 631–646.

Baker, L. E. The pupillary response conditioned to subliminal auditory stimuli. *Psychol. Monogr.*, 1938, **50**(3, Whole No. 223).

Binder, A. A statistical model for the process of visual recognition. *Psychol. Rev.*, 1955, **62**, 119–129.

Blackwell, H. R. Psychophysical thresholds: Experimental studies of methods of measurement. *Engng. res. Bull., U. Mich.*, 1953, No. 36.

Bressler, J. Illusion in the case of subliminal visual stimulation. *J. gen. Psychol.*, 1931, **5**, 244–250.

Cason, H., & Katcher, N. An attempt to condition breathing and eyelid responses to a subliminal electric stimulus. *J. exp. Psychol.*, 1933, **16**, 831–842.

Chatterjee, B. B., & Eriksen, C. W. Conditioning and generalization of GSR as a function of awareness. *J. abnorm. soc. Psychol.*, 1960, **60**, 396–403.

Coover, J. E. Experiments in psychical research. *Psychical Res. Monogr.*, 1917, No. 1.

Davis, R. C. Motor responses to auditory stimuli above and below threshold. *J. exp. Psychol.*, 1950, **40**, 107–120.

Diven, K. Certain determinants in the conditioning of anxiety reactions. *J. Psychol.*, 1937, **3**, 291–308.

Dixon, N. F. The effect of subliminal stimulation upon autonomic and verbal behavior. *J. abnorm. soc. Psychol.*, 1958, **57**, 29–36.

Dollard, J., & Miller, N. E. *Personality and psychotherapy.* New York: McGraw-Hill, 1950.

Dulany, D. E. Hypotheses and habits in verbal "operant conditioning." *Psychol. Rev.*, in press.

Dulany, D. E., & Eriksen, C. W. Accuracy of brightness discrimination as measured by concurrent verbal responses and GSRs. *J. abnorm. soc. Psychol.*, 1959, **59**, 418–423.

Dunlap, K. The effect of imperceptible shadows on the judgment of distance. *Psychol. Rev.*, 1900, **7**, 435–453.

Eriksen, C. W. Subception: Fact or artifact? *Psychol. Rev.*, 1956, **63**, 74–80. (a)

Eriksen, C. W. An experimental analysis of subception. *Amer. J. Psychol.*, 1956, **69**, 625–634. (b)

Eriksen, C. W. Prediction from and interaction among multiple concurrent discriminative responses. *J. exp. Psychol.*, 1957, **53**, 353–359.

Eriksen, C. W. Discrimination measured by multiple concurrent responses. *Psychol. Rep.*, 1959, **5**, 741–750.

Eriksen, C. W., Azuma, H., & Hicks, R. B. Verbal discrimination of pleasant and unpleasant stimuli prior to specific identification. *J. abnorm. soc. Psychol.*, 1959, **59**, 114–119.

Fisher, C. Dreams and perception: The role of preconscious and primary modes of perception in dream formation. *J. Amer. Psychoanal. Ass.*, 1954, **2**, 389–445.

Freud, S. *Collected Papers.* Vol. IV. London: Hogarth, 1925.

Fuhrer, M. J., & Eriksen, C. W. The unconscious perception of the meaning of verbal stimuli. *J. abnorm. soc. Psychol.*, 1960, in press.

Fullerton, G. S., & Cattell, J. McK. On the perception of small differences. *Phil. Ser.*, 1892, No. 2.

Garner, W. R., Hake, H. W., & Eriksen, C. W. Operationism and the concept of perception. *Psychol. Rev.*, 1956, **63**, 149–159.

Goldiamond, I. Indicators of perception: I. Subliminal perceptions, subception unconscious perception: An analysis in terms of psychophysical indicator methodology. *Psychol. Bull.*, 1958, **55**, 373–412.

Hilgard, E. R., Dutton, C. E., & Helmick, J. S. Attempted pupillary conditioning at four stimulus intervals. *J. exp. Psychol.*, 1949, **39**, 683–689.

Hilgard, E. R., Miller, J., & Ohlson, J. A. Three attempts to secure pupillary conditioning to auditory stimuli near the absolute threshold. *J. exp. Psychol.*, 1941, **29**, 89–103.

Kolers, P. A. Subliminal stimulation in problem-solving. *Amer. J. Psychol.*, 1957, **70**, 437–441.

Krasner, L. Studies of the conditioning of verbal behavior. *Psychol. Bull.*, 1958, **55**, 148–170.

Krieckhaus, E. E., & Eriksen, C. W. A study of awareness and its effects on learning and generalization. *J. Pers.*, 1960, in press.

Lacey, J. I., & Smith, R. L. Conditioning and generalization of unconscious anxiety. *Science*, 1954, **120**, 1045–1052.

LAZARUS, R. S. Subception: Fact or artifact? A reply to Eriksen. *Psychol. Rev.*, 1956, **63**, 343–347.

LAZARUS, R. S., & MCCLEARY, R. A. Autonomic discrimination without awareness: A study of subception. *Psychol. Rev.*, 1951, **58**, 113–122.

MALAMUD, W., & LINDER, F. E. Dreams and their relationship to recent impressions. *Arch. Neurol. Psychiat.*, 1931, **25**, 1081–1099.

METZNER, C. A., & BAKER, L. E. The pupillary response conditioned to subliminal auditory stimuli: A control experiment. *Psychol. Bull.*, 1939, **36**, 625. (Abstract)

MILLER, J. G. Discrimination without awareness. *Amer. J. Psychol.*, 1939, **52**, 562–578.

NEWHALL, S. M., & SEARS, R. R. Conditioning finger retraction to visual stimuli near the absolute threshold. *Comp. psychol. Monogr.*, 1933, **9**, No. 43.

OSGOOD, C. E. Motivational dynamics of language behavior. In M. R. Jones (Ed.), *Nebraska symposium on motivation:* Univer. Nebraska Press, 1957.

PEIRCE, C. S., & JASTROW, J. On small differences of sensation. *Mem. Nat. Acad. Sci.*, 1885, **3**, 73–83.

PERKY, C. W. An experimental study of imagination. *Amer. J. Psychol.*, 1910, **21**, 422–452.

POETZL, O. Experimentell erregte Traumbilder in ihren Beziehungun zum indirekten. *Sehen. Z. ges. Neurol. Psychiat.*, 1917, **37**, 278–349.

SHEVRIN, H., & LUBORSKY, L. The measurement of preconscious perception in dreams and images: An investigation of the Poetzl phenomena. *J. abnorm. soc. Psychol.*, 1958, **56**, 285–294.

SMITH, G. J. W., & HENRIKSSON, M. The effect on an established percept of a perceptual process beyond awareness. *Acta psychol., Amsterdam*, 1955, **11**, 346–355.

SMITH, G. J. W., SPENCE, D. P., & KLEIN, G. S. Subliminal effects of verbal stimuli. *J. abnorm. soc. Psychol.*, 1959, **59**, 167–176.

STECKLE, L. C. Two additional attempts to condition the pupillary reflex. *J. gen. Psychol.*, 1936, **15**, 369–377.

STECKLE, L. C., & RENSHAW, S. An investigation of the conditioned iridic reflex. *J. gen. Psychol.*, 1934, **11**, 3–23.

STERN, F. An investigation of pupillary conditioning. Unpublished doctoral dissertation, Univer. Washington, 1948.

TAYLOR, F. W. R. The discrimination of subliminal visual stimuli. *Canad. J. Psychol.*, 1953, **7**, 12–20.

TITCHNER, E. B., & PYLE, W. H. The effect of imperceptible shadows on the judgment of distance. *Proc. Amer. Phil. Soc.*, 1907, **46**, 94–109.

VINACKE, W. E. Discrimination of color and form at levels of illumination below conscious awareness. *Arch. Psychol.*, 1942, **38**, 15–16.

WEDELL, C. H., TAYLOR, F. V., & SKOLNICK, A. An attempt to condition the pupillary response. *J. exp. Psychol.*, 1940, **27**, 517–531.

WILCOTT, R. C. A search for subthreshold conditioning at four different auditory frequencies. *J. exp. Psychol.*, 1953, **46**, 271–277.

WILLIAMS, A. C., JR. Perception of subliminal visual stimuli. *J. Psychol.*, 1938, **6**, 187–199.

(Received January 8, 1960)

Psychological Review
1960, Vol. 67, No. 5, 301–316

SOME IMPLICATIONS OF A SOCIAL LEARNING THEORY FOR THE PREDICTION OF GOAL DIRECTED BEHAVIOR FROM TESTING PROCEDURES[1]

JULIAN B. ROTTER

Ohio State University

Many sophisticated observers are aware that a wide gap exists between personality theory and the techniques or procedures used to measure personality variables. The low level of prediction of such testing procedures may well be a function of the failure to apply the theories themselves to the methods of measurement. Particularly, it is a failure to apply an analysis of the determinants of behavior in general to the specific test taking behavior of the subjects (*S*s).

The gap itself may be described as having three aspects. The first of these relates to the question of the constructs used in the theory and the constructs which the tests were developed to measure. In many instances rather than devising tests which measure specific theoretical constructs which are carefully defined and for which the test behavior can be understood as a logical referent, the descriptive constructs used to classify test response do not logically relate to the new theoretical constructs but are bent or twisted to measure the new variables. That is, test constructs which were used to classify test responses developed earlier are "translated" to be measures of the new variables. Examples of this are use of Rorschach variables such as color, movement, and shading which arose from imagery-type theory, to measure such constructs as "ego strength," "rigidity," and "tolerance for ambiguity." The Rorschach was not developed to assess such variables and in translating or twisting older methods of Rorschach scoring to measure these variables it is quite likely that a loss of prediction results.

A second aspect of this gap between personality theory and methods of measurement of personality lies in the testing procedure itself. For example, where the theory may emphasize the significance of differences in behavior in the presence of authority figures vs. peers or males vs. females, the formal test procedure assumes no such variables are important. That is, no difference in interpretation of test results follows from the fact that the examiner may have a different social stimulus value in one case than in another or under one set of conditions rather than another. In such an instance although the theory itself recognizes (and experimental data such as Gibby, Miller, & Walker, 1953, and Lord, 1950, support) the importance of the effect on behavior of the nature of the social stimulus, the test procedure itself does not take it into account. An example would be in the application of Murray's theory (1952) which sees behavior as a function of internal *needs* and environmental *presses*. Tests have been developed using this theory (Thematic Apperception Test, as clinically used; Edwards Personal Preference Schedule) which presume to measure the strength of various needs but fail to account for the test behavior as a function of the testing situation itself (an environmental press) as one of the variables determining the test behavior. Other characteristics of this discrepancy between theory and test taking procedure will be discussed more fully later.

---

[1] I am indebted to Shephard Liverant for his helpful comments and suggestions about this paper.

A third aspect of the gap lies in the area of inference from test behavior. The issue here is that there is an absence of logic or contradiction in the assumed relationship between what the *S* does, or test behavior, and what is inferred from such behavior. Peak (1953) and Butler (1954) among others have discussed this problem earlier. Jessor and Hammond (1957) have noted such a gap in the inferences made from the Taylor Anxiety Scale. Another example could be drawn from the Edwards Personal Preference Schedule (Edwards, 1953) in which *S*s are asked to state their preferences for different kinds of goals but there is no theoretical basis provided to allow one to make predictions about *nontest behavior* from such preferences. Of course, it can be assumed that the preferences have some one-to-one relationship with some criterion behavior, but it is unlikely that even the test authors would make such a theoretical commitment. In other words, it is not clear exactly what can be predicted or should be predicted from the test responses. Individuals using such tests, however, can defend themselves by stating that prediction is after all an empirical matter and one has to find out what can or should be predicted. It is likely, however, that the construction of tests which are systematically or theoretically pure, in that they are devised to measure specific variables or to make specific predictions, with the method of measurement and inference consistent with the theory will ultimately provide much better predictions of behavior as well as a test of the utility of the theory itself.

The purpose of this paper is to explicate some of the implications of a social learning theory of personality for the measurement of personality variables. The particular point of emphasis is the measurement of goal directed behavior conceptualized in social learning terms as *need potential*. Secondarily, the paper aims at illustrating the nature of the relationship between testing procedures and inference about behavior more generally.

In social learning theory (Rotter, 1954) the basic formula for the prediction of goal directed behavior is as given below:

$$\mathrm{BP}_{x,\mathrm{s}_1,\mathrm{R}_a} = \mathrm{f}(\mathrm{E}_{x,\mathrm{R}_a,\mathrm{s}_1} \mathbin{\&} \mathrm{RV}_{a,\mathrm{s}_1}) \qquad [1]$$

The formula may be read as follows: The potential for behavior *x* to occur in Situation 1 in relation to Reinforcement *a* is a function of the expectancy of the occurrence of Reinforcement *a* following Behavior *x* in Situation 1 and the value of Reinforcement *a,* in Situation 1. Such a formula, however, is extremely limited in application for it deals only with the potential for a given behavior to occur in relationship to a single specific reinforcement. The prediction of responses from personality tests requires a more generalized concept of behavior and the formula for these broader concepts is given below:

$$\begin{aligned}\mathrm{BP}_{(x-n),\mathrm{s}_{(1-n)},\mathrm{R}_{(a-n)}} &= \mathrm{f}(\mathrm{E}_{(x-n),\mathrm{s}_{(1-n)},\mathrm{R}_{(a-n)}} \\ &\qquad \mathbin{\&} \mathrm{RV}_{(a-n),\mathrm{s}_{(1-n)}}) \qquad [2]\end{aligned}$$

This may be read: The potentiality of the functionally related Behaviors *x* to *n* to occur in the specified Situations 1 to *n* in relation to potential Reinforcements *a* to *n* is a function of the expectancies of these behaviors leading to these reinforcements in these situations and the values of these reinforcements in these situations. For purposes of simplicity of communication, the three basic terms in this formula have been typically referred to as need potential, freedom of movement, and need value as in the third formula below:

$$NP = f(FM \& NV) \qquad [3]$$

In this formula the fourth concept, that of the psychological situation, is implicit. The variables referred to above and operations for measurement have been defined and further explicated in a previous publication (Rotter, 1954).

In order to illustrate the social learning theory implications for measurement of personality and for the measurement of goal directed behavior, it seems expedient to consider three basic approaches to this problem based on the number of determinants used theoretically to predict such goal directed behavior and the problems, limitations, and advantages of each approach.

STRENGTH OF NEED AS A BASIS FOR PREDICTING BEHAVIOR

Although many esoteric systems of prediction utilize essentially the strength of need, drive, or instinct approach, this method can be described as the simplest or least complicated approach. Basically a series of constructs are formulated more often on an a priori basis than empirically, or at least on a presumed clinical rather than experimental basis. These descriptive terms may refer to instincts, drives, needs, factors, entities, or vectors of the mind (i.e., the Minnesota Multiphasic Personality Inventory, Edwards Personal Preference Schedule, Rorschach, Humm-Wadsworth, etc.). They all have in common that there is more than one basic characteristic, that these two or more characteristics are in some way measurable along a continuum and presumably the individual's behavior can be predicted from the characteristics which are "stronger" and the characteristics which are "weaker."

Sometimes the personality disposition can be predicted from the strength of other constructs according to either simple or complex statements of relationship formally postulated, hypothesized, or informally asserted. These relationships can become quite complex as in psychoanalysis or quite esoteric as in Szondi's explanation that motivated behavior is a result of the interaction of dominant and recessive genes. Because the methods of measurement in some instances cannot be direct, as in the assessment of unconscious drives in psychoanalysis, an impression of great complexity is given but regarded entirely from the point of view of the prediction of behavior, the system may still have a simple character. The potential for a given kind of behavior is still directly predictable from the strength of the drive, instincts, needs, or energies postulated.

There is another form of this model in which the various drives, dispositions, or needs are considered to interact. For example, the individual may be conceived of as being controlled by his intellect and his emotions, but his behavior must be understood in light of the interactions of these two forces with a third variable, the will as in the Rorschach Test (1942). Again, this makes complex the calculation of strength or weakness but does not change the overall method of making predictions. Whether dealing with will, intellect and emotions or ego, superego and id, the effect of interaction is only to increase or decrease the tendency of one of the needs to function or to strengthen or weaken one of them or perhaps to produce a fourth or fifth additional need. The basic method of prediction stays the same although the calculation of strength or weakness in such a model becomes more difficult.

The obvious problem, of course, for such rudimentary method of prediction is how to predict anything at all. If a system included five instincts or needs and these are ordered on some metric system from high to low, does one assume that the person will always act in the fashion to be predicted from his strongest or highest need? If an *S* is more oral than anal, does he always act in an oral fashion? Actually the most logical assumption in regard to any specific instance is that he will always act the same way. One might presume on a statistical basis, as it is sometimes done, if the individual is at the 70th percentile on Need A and at the 30th percentile on Need B, 70% of the time he would act in one fashion and 30% of the time in the other over some undefined period of time. However, the only sensible statistical or logical prediction in any specific instance, if no other variables are concerned, is that he would act in accordance with the higher need. This might still give fairly good prediction if only 2 variables are involved, but if 20 variables are involved and many of them are very close in value or "strength," then the amount of error begins to increase. In fact, it becomes a problem to predict even slightly above chance and, indeed, except for some limited and highly controlled experimental situations, this is the problem in psychology now. A recent illustration of this failure is reported in a carefully controlled study by Little and Shneidman (1959) who failed to find much relationship between interpretations of psychological tests (Rorschach, MAPS, TAT, and MMPI) and anamnestic data. Loevinger (1959) summarizing the predictiveness of individual tests in the recent *Annual Review of Psychology* states, "To date the only tests which meet standards for individual prediction are those of general ability" (p. 305). Previous reviewers have made similar statements.

Another problem which arises in the prediction of behavior with this simple model is that it soon becomes apparent that the strength of a wish, need, or drive to achieve some goal such as being taken care of, obtaining love, or injuring someone is not a good predictor of the occurrence of behavior directed towards the achieving of that goal. To some extent this problem can be dealt with by the notion of interaction of needs, but usually in order to account for the discrepancy between need or wish and behavior, constructs of the same order do not provide sufficient explanatory basis. It is actually necessary to postulate some other kinds of internal constructs to account for the discrepancy between what might be called wish, desire, need, or instinct and observable behavior.

In the measurement of these need strengths all varieties of tests and devices have been used. To some extent, the personality questionnaire is utilized a little more by people adopting such a predictive scheme as that described above, but also projective tests, observation, interviewing, and many other techniques of personality measurement have been used in this fashion. Test construction methodologies may currently be more sophisticated in that they control for social desirability of items, motivation, faking, lying, and inability to understand directions. Recent tests may also rely on purification of factors, cross-validation, or item analysis. However, with all these "modern improvements" in test design one is still left with a series of figures which are of doubtful utility for the actual prediction of behavior at a level satisfactory for either practical application or for the clarification of theoretical issues.

THE ADDITION OF AN EXPECTANCY CONSTRUCT IN THE PREDICTION OF GOAL DIRECTED BEHAVIOR

The absence of additional variables explicitly defining the relationship of need and behavior appears to be not so much a matter of simple theoretical structure as it is merely the absence of any real explicit theory about human behavior. The development of a predictive model which recognizes the discrepancy between need and behavior and tries to systematically take it into account represents a second level of sophistication.

At an earlier date perhaps psychoanalysis dealt with this problem most effectively in introducing concepts such as repression, sublimation, suppression, defense, reaction formation, etc., to account for the discrepancy between observed behavior and the presumed internal drive, need, or instinctual urge.

At this level of theorizing some systematic variable is added to the internal motivational state in order to predict behavior. Perhaps another way of saying this is that in addition to some measure of preference or value of a specific goal another systematic concept must be introduced, not only in a hit-or-miss fashion but perhaps directly into our assessment procedure. The psychoanalytic solution has been criticized because many specific concepts are introduced to account for a discrepancy between drives, urges, or needs and observable behavior, but these concepts are not readily measurable. In addition, one does not know when one explanation, i.e., reaction formation, is the explanatory concept or another such as sublimation or simple repression.

In social learning theory (Rotter, 1954) it is presumed that the relationship between goal preference (reinforcement value) and behavior can be determined only by introducing the concept of the individual's expectancy, on the basis of past history, that the given behavior will actually lead to a satisfying outcome rather than to punishment, failure, or, more generally, to negative reinforcement. Since the early formulations of Tolman (1932), expectancy theories have become more and more widely relied upon both in human learning and personality theories. It is possible to conceptualize more specific constructs such as repression and reaction formation as only special cases of an expectancy for severe punishment and that a more general relationship holds which includes perhaps all of these and also expectancies for punishment or failure of which the individual is quite aware. For example, an individual may wish very much to be a good dancer and to dance with members of the opposite sex. He makes no attempts, however, to dance at a party or a dance because he can tell you "but I look like a fool when I go out on the dance floor." We need, in other words, to introduce no specific construct involving the "unconscious" to explain the discrepancy between his wish and his behavior. The *S* may or may not be aware of expectancies which influence his behavior. Whether or not he is aware of them may affect the degree to which these expectancies change with new experience as well as other variables. The degree of awareness may be an important additional variable; however, the level of expectancy itself is the broader variable which bears a direct relationship to the potential occurrence of a specified behavior.

The question arises, then, of how one takes into account such factors in an actual testing situation. It could be said that no one is really so naive as to believe that the strength of an internal motivational condition or need is a direct predictor of behavior. Somehow or other, whether or not the individual had learned a given behavior or expected it to work is also an important aspect of prediction. However, more often than not this aspect of prediction has been treated as a source of error, something to be eliminated if possible, both in testing or in the validation of a test instrument. As a matter of fact, many currently used instruments attempting to assess the strength of motives, drives, or needs are usually confounded. Although they may be quite sophisticated in methodology, the test items or the test variables usually refer in part to what the individual did, i.e., overt behavior ("I frequently lose my temper"), in part to what he wished ("I would like to have more friends"), and in part to what he expected to be the outcome of his own behavior ("I feel that other people do not appreciate my good intentions"). To some extent these impure items probably add to prediction by providing more than one kind of referent for behavior, but the nonsystematic way in which they are used also limits prediction.

In trying to predict goal directed behavior from tests, two possibilities are open. One of these is to attempt to predict behavior from other behavior which presumably is functionally or predictively related to the test behavior. What this involves is analyzing test situations as behavioral samples under a given set of test conditions. For example, to assess dependent behavior one could use direct observation techniques, perhaps in problem solving situations, in which the *S* is scored for help-seeking behavior (cf. Naylor, 1955). In questionnaires the items should refer to what the *S* does, not to what he expects, wishes, or feels. The use of behavior samples for predictions or the regarding of all kinds of tests including projective tests essentially as samples of behavior to be analyzed in terms of what the *S* does under these conditions has been described elsewhere (Rotter, 1954). Like the work sample test in industry it undoubtedly provides the best prediction to a limited specific behavioral criterion since it requires the fewest intermediate constructs and the fewest assumptions regarding the action of other variables.

However, there are many problems, both theoretical and clinical, when it is important to break down this behavioral measure into its major determinants of reinforcement value and expectancy for the occurrence of the reinforcement. For example, in psychotherapy an understanding of how some behavior or group of behaviors may be most readily changed requires analysis into at least these two components. Even when strictly concerned with predictions of behavior in a broad band of life situations, rather than change, analysis into separate determinants may provide greater prediction than a work sample or behavioral technique. In this second alternative it is important either to control or systematically vary the other variable or measure both. For example, Liverant (1958) has measured some needs by presenting pairs of items involving goal preference matched for social desirability, and Jessor and Mandell (Mandell, 1959) are developing a similar test to measure expectancy for success in satisfying the same needs.

In using projective material such as the TAT, Crandall (1951) has demonstrated that expectancy for need satisfaction, for which the term freedom of movement is used in social learning theory, can be reliably measured by selecting particular kinds of referents. The work of Mussen and Naylor (1954), Kagan (1956), and Lesser (1957) gives strong evidence that the relationship between theme counts of aggression on the TAT and overt aggressive behavior depends to a large extent on whether or not that overt behavior is socially acceptable in the *S*s' own homes or social climates. The relationship between theme count and overt aggressive behavior appears to hold only when the *S*s do not have a high expectancy that aggressive behavior will be punished. Atkinson and Reitman (1956) report that in a number of studies of need achievement, it has become clear that prediction of behavior is enhanced if, in addition to taking a measure of need achievement based upon achievement theme counts in TAT-like material, an additional measure of expectancy for success is also taken into account.

In dealing with this type of testing material the recently published study of Fitzgerald (1958) provides a more systematic analysis. Using a highly reliable sociometric technique of nomination of fraternity brothers as his behavioral criteria and dealing with the need dependency, Fitzgerald found no relationship between theme counts and overt behavior. Presumably, dependent behavior is not socially acceptable among male college students. He had, however, independent interview ratings of need value, that is preference or desire for dependency satisfactions and of freedom of movement, or expectancy that behavior directed toward achieving dependency would be satisfied.

He found that by using these measures he did obtain a significant correlation between theme counts and the *discrepancy* between need value and freedom of movement. More specifically, what he called a conflict score or score indicating the degree to which the individual preferred dependency or desired dependency satisfactions but expected that he could not achieve them correlated with theme counts for dependency.[2] On the other hand, an Incomplete Sentences Blank measure of dependency which utilized behavioral referents as well as reinforcement value and expectancy referents did show a low but significant relationship of the number of completions dealing with dependency with both the sociometric and interview measures of need potential or actual dependent behavior. Although an actual analysis was not made, it seems very likely that the reason for the correlation in the case of the ISB and not the TAT is that at least some of the ISB completions were descriptions of actual behavior. Possibly a purer measure of behavior would have shown a greater relationship to the sociometric and interview rating assessment of actual dependent behavior in life situations.

Should we build two instruments or at least two sets of testing operations to separately assess need value and freedom of movement, or should we attempt to use behavioral measures in order to make our predictions about behavior, we would still be faced with the problem of predicting in a specific situation. Given measures of six behavior potentials, however arrived at, the problem remains that of knowing which of these is likely to be the behavior preferred in some specific situation. One is again forced to predict that the behavior with the highest potential always occurs and one is limited again in prediction to a very low level of accuracy. In the laboratory situation where we can reduce the possible alternatives to two, significant, although not predictive results, are possible. In the life situation where the alternatives are very frequently of a large order, the question arises of whether or not any useful prediction is possible. This leads us to a third level of sophistication, one in which the psychological situation is one of the variables on which prediction is based.

---

[2] Whether or not the relationship between theme count and high reinforcement value and low expectancy is general is not yet known. It appears at this time to possibly depend on whether or not the test material and testing situation is conducive to the free expression of fantasy.

THE PSYCHOLOGICAL SITUATION AS A THIRD DETERMINANT OF GOAL DIRECTED BEHAVIOR

Few would deny that the psychological situation will affect the potential of occurrence of any behavior or class of behaviors. However, the fact that behavior will vary from situation to situation is most often treated as a source of error, something to be avoided. If possible, one should construct tests or find personality variables which rise somehow above the situation. It is probably no exaggeration to say that thousands of hours of wasted work have been done by psychologists in the vain goal of finding either tests or variables which would, somehow or other, predict regardless of the situation in which the test is given or regardless of the situation in which the predicted behavior is expected to occur.

There are three separate problems here which will be discussed as one basic problem. The first problem is to understand the effect of the testing situation on test results. For example, Sarason (1950) has provided an excellent discussion of some of the influential situational variables in intelligence testing. The second problem is to understand the nature of the criterion situation which affects the criterion measures. The third and ultimately most important problem is to devise our tests with full consideration of the nature of the test situation in order to predict behavior in other situations for which the test was constructed. In other words, we need to devise tests not to predict personality or needs or behavior in the abstract but in specified situations or classes of situations if we want high prediction. *We need to know and take into account the dimensions of situation similarity in devising test procedures.*

Cronbach (1956) has criticized the failure to regard differentially the criterion situations in which tests are applied. In regard to the test situation we have only attempted to standardize the test procedure but usually have ignored the importance of the social context in which the test is given. Perhaps the most important thesis of this paper is that the psychological situation needs to be understood and systematically considered in our predictive formula, not treated as a source of error or something that can be ignored because part of the total situation is standardized.

Recently there have been a number of studies which demonstrate that almost all tests are subject to faking, to instructional variation, to examiner influences, to testing conditions, etc., regardless of the type of test (Borislow, 1958; Green, 1951; Gross, 1959; Mussen & Scodel, 1955; Rotter, 1955). The general inference drawn from these studies is that the tests are poor. Actually the implication of such findings is that we are making inefficient use of our tests. If the test situation for many personality tests is one in which social conformity or acceptability is easily achieved and no other satisfactions are given up in achieving acceptability, then for some purposes this motive should be controlled. However, the test situation can also be utilized to measure the importance of social conformity for the individual. *What we call faking is only our recognition of the fact that the S is taking the test with a different purpose or goal than the one the examiner wants him to have.* For some purposes it might be important to understand what kind of goals he exhibits in this kind of situation. More often than not we simply try to control what we should be studying. For example, in giving intelligence tests it might be better to study systematically the effect on performance of encouragement and discouragement rather than to attempt some mythical neutral attitude which is presumably the same for every examiner. Knowledge of the effects of situational variations would be of particular value in understanding the frequently diverse and contradictory results of apparently similar research investigations. For example, Henry and Rotter (1956) found that large, predicted differences were obtained between two comparable groups on the Rorschach test if one group was reminded before the regular instructions that the test had been used frequently to study psychopathology. An obvious implication of this study is that investigators using this same test in the college laboratory and in the clinical setting may well produce diverse results.

Another example of how the situation can be used in testing is provided by the patient who is being assessed for possible benefit from psychotherapy. If the clinic or hospital can provide both male and female therapists and also therapists who rely on support and direction as well as therapists who remain distant and passive, then the testing procedures can be varied so that those situational influences are present. The testing could provide information to indicate what kind of therapy and what kind of therapist is likely to provide *this* patient with the most efficient conditions for relearning. For more conventional purposes of clinical testing, it is still more important to know under what conditions the patient behaves in a paranoid fashion and under what conditions he does not, than it is to know how many percentiles of paranoia he has.

Not only can the test situation itself be analyzed as a behavioral sample but situational referents can be incorporated into the content of items by systematic sampling. For example, questionnaire items can deal with the *conditions* under which the *S* feels tense, nervous, happy, has headaches, etc. as Mandler and Sarason (1952) have done for some intellectual test taking situations. Similarly, projective methods, particularly TAT-type tests, can systematically vary the situation through the selection of test stimuli as has been done by Crandall (1951) and McClelland, Atkinson, Clark, and Lowell (1953). More recently Murstein (1959) has suggested a conceptual model for stimulus variation with thematic techniques.

The many studies indicating marked effects of testing conditions suggest that it is of great importance in the publication of any test that descriptions of the differences in test results that are likely to be associated with different kinds of testing situations be provided. No test can be adequately understood unless the data regarding its standardization or use includes systematic descriptions of the differences in test results which are a function of different kinds of testing conditions and different kinds of purposes in taking the test for similar samples of *S*s. Only when we know whether an *S* is likely to produce different test results when he is taking the test to demonstrate how imaginative he is as compared to taking it to prove that he needs help will we be adequately able to understand the meaning of test results and to predict future behavior from them.

There have been personality theorists who have made much of the importance of the individual's life space. Kantor (1924) was one of the first to emphasize that the basic datum of psychology is the interaction of an individual and his meaningful environment. For Kantor, people do not have internal characteristics in the same sense as for other theorists; rather they have a reactional biography of interactions with the environment. Lewin (1951) has also emphasized the importance of the life space or psychological situation in the determination of human behavior. Brunswik (1947) has repeatedly called for analyses of and sampling of psychological situations for predictive purposes. Helson (1948) has applied his theory of Adaptation Level to social psychology stating that the effect of the total field can be quantified by careful ordering of the field of exposed stimuli. Recent concern with the importance and need for systematic study of situation variation has been expressed by Allport (1958) and Cronbach (1957).

In a more limited and less systematic way, psychoanalysis has suggested, in a few areas, that certain kinds of goal directed behavior depended upon the psychological situation. This is done in making distinctions between the individual's potential response to authority figures vs. nonauthority figures and males vs. females. Beyond this, little systematic analysis of differences in life situations has been made by the traditional analyst. Murray's (1952) formulation of the nature of personality stresses that behavior is a function of the interaction of an individual with a psychological situation which he felt could be categorized as "press." At a more specific level Atkinson and Raphelson (1956) have shown the value of including situational variables in studying achievement behavior. This general point of view has also been represented in sociology by Thomas (1951) and Coutu (1949).

In social learning theory, it has been hypothesized that the situation operates primarily by providing cues for the *S* which are related to the magnitude of his expectancies for reinforcement for different behaviors. The effect on the value of the reinforcement itself operates through expectancies for associated or subsequent reinforcements which may differ from situation to situation. It has also been hypothesized that situations may be usefully categorized in terms of the predominant reinforcements as culturally determined for any large or small culture group. There are many other possible ways of categorizing situations depending upon the predictive purposes involved.[3] Methods of determining generality or determining the dimensions of similarity among situations have been described in an earlier paper (Rotter, 1955).

---

[3] Several writers have pointed out the difficulty of identifying situations independently of behavior. That is, how can one describe a situation as one would a physical stimulus independently of the *S*'s response? The problem is not different from that of describing stimuli along dimensions of color although perhaps vastly complicated in social or other complex situations. In the case of color stimuli ultimately the criteria is a response of the scientist or observer, sometimes a response to an intermediate instrument, and one that is at the level of sensory discrimination and so leads to high observer agreement. In the case of the social situation, the level of discrimination is common sense based on an understanding of the culture rather than the reading of an instrument. As such, reliability may be limited but still be sufficiently high to considerably increase prediction. In this way specific situations could be identified as school situations, employment situations, girl friend situations, etc. For the purpose of generality various kinds of psychological constructs could be devised to arrive at classes of situations which have similar meaning to the *S*. The utility of such classes would have to be empirically determined depending on the *S*'s response. The objective referents for these situations, which provide the basis for prediction, however, can be independent of the specific *S*. That is, they can be reliably identified by cultural, common sense terms.

Two illustrative studies of an increasing number of experimental analyses of behavior which vary both internal characteristics and the psychological situation systematically in the same study are described below. These studies follow the basic paradigm that the presumed relevant individual (personality) and situational (experimental manipulations) variables can be observed simultaneously and their interaction studied.

ILLUSTRATIVE STUDIES VARYING BOTH THE SITUATION AND INTERNAL CHARACTERISTICS

A recent doctoral dissertation by James (1957) illuminates very clearly the potential of greater prediction when both the situation and the internal characteristic are varied in the same study. The behavior being studied by James involved a variety of learning variables, including acquisition, changes or shifts, extinction, generalization, and recovery of verbalized expectancies for gratification. Two general hypotheses were involved in this study growing out of previous work by Lasko (1952), Phares (1957), Neff (1956), and James and Rotter (1958). Hunt and Schroder (1958) have also dealt with what appears to be a related variable. The first of these hypotheses might be stated as a situational one. That is, that the nature of a learning process differs depending upon whether or not the situation is one in which the reinforcements that occur are a direct outcome of some internal characteristic of the individual such as skill, a physical characteristic, or whatever, versus a situation where the reinforcements are essentially controlled by someone else or by chance or by conditions or powers beyond the *S*'s control. Perhaps a good example of the latter would be a dice game or the winning of a door prize or having soup spilled on one because a waiter tripped, etc. James utilized line and angle matching tasks reinforcing each *S* positively on his guesses in six of the eight training trials. He specifically hypothesized when the situation is structured in such a way that the *S* expects the occurrence of reinforcements to be beyond his control or partly beyond his control, increments and decrements in expectancy for gratification as a result of experience are smaller, the number of unusual shifts, that is, shifts up after failure or down after success, are greater, extinction is faster, and there is less generalization from one task to another and greater recovery following extinction.

The measurement of individual differences in this study followed from the previous work of Phares which suggested that individuals can be differentiated in the degree to which they see the world and the things that happen to them as controlled by others or as determined by chance or unpredictable forces. The second hypothesis, then, was that all the differences which would occur as a result of the situational conditions would also be true of individuals within all groups as a function of their general attitude towards "control of reinforcement."

In order to predict the individual differences in attitude, James enlarged and revised the questionnaire first devised by Phares. This was given to all *S*s at the end of each experiment. The results are most striking. All of the predicted outcomes hypothesized above resulting from the differences in instructions or situations were obtained and all were statistically significant with the exception of recovery following extinction, which showed a strong trend in the predicted direction. Similarly, within each group the individuals high as compared to low on the questionnaire differed significantly in exactly the same way as did the groups themselves as a result of the different instructions or situations presented. Although individual prediction was not the concern of this investigation, it is quite clear that a simple formula could be devised which could predict all of the learning variables involved in this study with a fair degree of accuracy. Certainly it is clear that a greater degree of accuracy is possible when both the situational and individual variables are taken into account. Perhaps far more important, this study indicates that various experimental paradigms in studying human learning are likely to produce different kinds of results. Whether or not a given learning task is one in which the *S* feels that success is dependent upon the experimenter's manipulation (for example, when he is expecting to predict a random sequence of red or green lights) or is the result of his own skill provides a crucial difference in the nature of the learning process itself.

The study of James, however, does not provide a satisfied feeling that it illustrates all of the problems of prediction involving both the individual's characteristics and the situation from which the prediction is made. It gives an almost too simple picture of the interaction of these two variables. Another recent dissertation by Moss (1958) suggests that this relationship can be more complex, and illustrates more clearly the effect of the testing situation on more commonly used types of assessment procedures.

Moss studied a general behavioral characteristic which he called cautiousness. Essentially this was defined as the avoidance of risk, the selection of the safest alternative in a situation where failure or negative reinforcement was possible. He varied the situation by reacting differently to three groups following the administration of a questionnaire which he described as a measure of social acceptability. One group was shown false norms at the conclusion of the questionnaire that indicated that they were in the ninetieth percentile of social acceptability for a college group. Another group of *S*s was shown that they were at the tenth percentile, and a third group was given no information about the results of this supposed test of social acceptability. He hypothesized that cautious behavior would increase with negative reinforcement. Immediately following this procedure, the *S*s were given two projective type tests and behavior on these tests was analyzed as to degree of cautiousness.

Prior to the giving of the "social acceptability" questionnaire the *S*s had been tested on the level of aspiration board. Behavior on the level of aspiration board (Rotter, 1942) was categorized into cautious or noncautious patterns.[4] The general tendency to seek safe alternatives in the obtaining of satisfactions then was measured in a situation in which the *S* himself has some control over failure or success.

One kind of behavior studied was that of sorting figures taken from the MAPS test. The *S* was presented the figures and asked to sort them into two piles any way he wished. The procedure was repeated a second time asking for a different kind of sort, and a third time. The sorts themselves were characterized as being safe or cautious in that they dealt with highly objective characteristics of these figures, or less safe in that they dealt with characteristics which were more abstract or had to be read into many of the figures. For example, sorts based on personality characteristics were considered as noncautious as opposed to safer or more cautious sorts such as those into groups of men and women, children and adults, Negroes and whites, etc.

---

[4] Cautious and noncautious behavior was characterized according to the patterns described by Rotter (1954, pp. 318–324). Patterns 1 and 3 were considered as noncautious and 2, 4, 7, and 8 as cautious patterns. The latter group are characterized by a variety of techniques presumably aimed at avoiding failure to reach explicit goals. Patterns 1 and 3 are characterized by higher expectancies than performance but within "normal" bounds and consequently a higher number of failures to reach one's estimates.

A second kind of behavior studied was the *S*'s response to a series of four TAT pictures. In this case the *S*'s stories were treated as Weisskopf (1950) has treated them with her measure of transcendence. A cautious or safe interpretation was one sticking close to the characteristics of the picture and one in which the theme itself was a common one.

Moss found some differences among his groups in the direction he had hypothesized, that is, that the threatened group, the group that was told that it was at the tenth percentile, showed greater cautiousness than the other two groups. The differences between groups, however, although consistently in the direction he predicted, only approached significance and were not large. However, when Moss divided his *S*s within groups into cautious and noncautious on the basis of their level of aspiration patterns, he found some highly significant differences. The cautious *S*s showed no significant differences among the three conditions. However, the noncautious *S*s showed significant differences between conditions. That is, noncautious *S*s in one condition responded differentially from noncautious *S*s in another condition. These differences were primarily due to greater noncautious behavior in the no-information group. Differences in test behavior between cautious and noncautious *S*s were also highly significant on both tests within the no-information group but not in the other groups.

In spite of the complexity of this study, a few findings seem relatively clear from an analysis of group means as well as significance of differences. *S*s who were cautious on the level of aspiration test, which is a somewhat free situation, were also cautious in the other test conditions. Of course, this does not mean that they were cautious in situations which were not perceived by them as evaluation situations. On the other hand, *S*s who were noncautious on the level of aspiration test appeared to maintain this greater risk taking behavior under test situations where no information about results was given. However, when they were negatively reinforced, they became more cautious and they also did not appear to be different from cautious *S*s under conditions where they were quite successful. Perhaps this is related to the presumed conservatism which follows from success. There was no consistent prediction from the level of aspiration situation to the two "projective tests" which could be made without considering the situation. In at least two of the situations the cautious *S*s were not significantly different from the noncautious *S*s. On the other hand, the psychological situation or the three different situations seemed to have no effect on the cautious *S*s. Only in the interaction of the noncautious *S*s with the situational variables was prediction possible from the level of aspiration test.

A similar type of result to that of Moss was recently reported by Lesser (1959). Lesser found that intercorrelations among various measures of aggression were significantly higher under experimental conditions of low anxiety than under conditions of high anxiety about aggression.

James' results suggest a rather simple relationship between dispositional and situational variables, but it is clear from the study of Moss and other studies that a simple additive or multiplicative relationship will not always describe the nature of the interaction. An important implication of this principle is that there is a general lack of efficiency in research studies in which only one set of variables, that is only dispositional or situational, are systematically varied, since the conclusions of the two sets of studies cannot be put together in a simple fashion. Unless both kinds of variables are systematically varied *within the same investigation,* both an understanding of the determinants of behavior and the prediction of it may suffer.

A striking example of the importance of studying the effects of dispositional and situational influences simultaneously is provided by Helson, Blake, Mouton, and Olmstead (1956). In applying Adaptation Level theory to a study of attitude change they found important interactional effects when situations were varied in external influence pressure and individuals were distributed on a measure of ascendancy–submissiveness.

It is true that many of the above propositions are obvious. Most psychologists recognize that there is a difference between overt behavior directed toward certain goals and the desires that individuals have to obtain these goals. Similarly, most psychologists know that the psychological situation is a determinant of the occurrence of a given behavior. The thesis here, however, is not merely that this is the case but that all of these variables must be ordered and studied systematically, in order to make predictions.

SUMMARY

The major contention of this paper has been that the prediction of goal directed behavior of human subjects from test procedures has been and will continue to be at an extremely low hit-or-miss level because of inadequate conceptualization of the problem. Findings are frequently not replicatable because of the failure to systematically differentiate behavior, reinforcement value, and expectancy as internal variables and to recognize that these variables are affected by the psychological situation.

The psychological situation of the patient in the clinic is so different from that of the elementary psychology student taking a test as part of an experiment that it is possible that the kinds of predictions which can be made in one situation would hardly hold in the other. The evidence that faking is possible and that different norms obtain when subjects are job applicants, employees, or volunteers does not necessarily mean that a test is no good. Nor is prediction essentially hopeless because it can be demonstrated that two experimenters, whether the same sex or opposite, or slight changes in the wording of instructions, will differentially affect test or experimental results. All of these things indicate only that the psychological situation, perhaps acting primarily through the expectancies they arouse by the cues present, considerably affect behavior. It is necessary that we do not consider such influence as error to be ignored, as difficulty to be avoided or as the problem of some other profession to investigate. Rather it is necesary to study these influences and consider them regularly and systematically in a predictive schema. That is, for some purposes, factors such as social desirability of items, examiner's behavior, and the subject's goals in the test situation should be controlled, and in other cases they should be allowed to vary. In all cases, however, they must be systematically considered.

Implicit in this entire paper is the belief that a satisfactory theory of goal directed behavior is a primary prerequisite for developing adequate tests. Knowledge of statistics and test construction procedures can be valuable but they cannot supplant an adequate theory of behavior which is applied to the test taking behavior itself.

To arrive at a fully systematic model for relating these general or high order constructs and to coordinate them in turn to lower level sets of content variables, devised for different purposes, will be a long and arduous but rewarding task.

REFERENCES

ALLPORT, G. W. What units shall we employ? In G. Lindzey (Ed.), *The assessment of human motives.* New York: Rinehart, 1958.

ATKINSON, J. W., & RAPHELSON, A. C. Individual differences in motivation and behavior in particular situation. *J. Pers.*, 1956, **24**, 349–363.

ATKINSON, J. W., & REITMAN, W. R. Performance as a function of motive strength and expectancy of goal attainment. *J. abnorm. soc. Psychol.*, 1956, **53**, 361–366.

BORISLOW, B. The Edwards Personal Preference Schedule and fakability. *J. appl. Psychol.*, 1958, **42**, 22–27.

BRUNSWIK, E. *Systematic and representative design of psychological experiments.* Berkeley: Univer. California Press, 1947.

BUTLER, J. M. The use of a psychological model in personality testing. *Educ. psychol. Measmt.*, 1954, **14**, 77–89.

COUTU, W. *Emergent human nature.* New York: Knopf, 1949.

CRANDALL, V. J. Induced frustration and punishment-reward expectancy in thematic apperception stories. *J. consult. Psychol.*, 1951, **15**, 400–404.

CRONBACH, L. J. Assessment of individual differences. In P. R. Farnsworth and Q. McNemar (Eds.), *Annu. Rev. Psychol.*, 1956. Stanford, Calif.: Annual Reviews, 1956. Pp. 173–196.

CRONBACH, L. J. The two disciplines of scientific psychology. *Amer. Psychologist*, 1957, **12**, 671–684.

EDWARDS, A. L. *Manual of the Edwards Personal Preference Schedule.* New York: Psychological Corp., 1953.

FITZGERALD, B. J. Some relationships among projective test, interview and sociometric measures of dependent behavior. *J. abnorm. soc. Psychol.*, 1958, **56**, 199–204.

GIBBY, R. G., MILLER, N. R., & WALKER, E. L. The examiner's influence on the Rorschach protocol. *J. consult. Psychol.*, 1953, **17**, 425–428.

GREEN, R. F. Does a selection situation induce testees to bias their answers on interest and temperament tests? *Educ. psychol. Measmt.*, 1951, **11**, 503–515.

GROSS, L. R. Effects of verbal and nonverbal reinforcement in the Rorschach. *J. consult. Psychol.*, 1959, **23**, 66–68.

HELSON, H. Adaptation level as a basis for quantitative theory of frames of reference. *Psychol. Rev.*, 1948, **55**, 297–313.

HELSON, H., BLAKE, R. R., MOUTON, JANE S., & OLMSTEAD, J. A. Attitudes as adjustments to stimulus, background, and residual factors. *J. abnorm. soc. Psychol.*, 1956, **52**, 314–322.

HENRY, EDITH M., & ROTTER, J. B. Situational influences on Rorschach responses. *J. consult. Psychol.*, 1956, **20**, 457–462.

HUNT, D. E., & SCHRODER, H. M. Assimilation, failure-avoidance, and anxiety. *J. consult. Psychol.*, 1958, **22**, 39–44.

JAMES, W. H. Internal versus external control of reinforcement as a basic variable in learning theory. Unpublished doctoral dissertation, Ohio State Univer., 1957.

JAMES, W. H., & ROTTER, J. B. Partial and one hundred percent reinforcement under chance and skill conditions. *J. exp. Psychol.*, 1958, **55**, 397–403.

JESSOR, R., & HAMMOND, K. R. Construct validity and the Taylor anxiety scale. *Psychol. Bull.*, 1957, **54**, 161–170.

KAGAN, J. The measurement of over aggression from fantasy. *J. abnorm. soc. Psychol.*, 1956, **52**, 390–393.

KANTOR, J. R. *Principles of psychology.* Vols. 1, 2. New York: Knopf, 1924.

LASKO, A. A. The development of expectancies under conditions of patterning and differential reinforcement. Unpublished doctoral dissertation, Ohio State Univer., 1952.

LESSER, G. S. The relationship between overt and fantasy aggression as a function of maternal response to aggression. *J. abnorm. soc. Psychol.*, 1957, **55**, 218–222.

LESSER, G. S. Population differences in construct validity. *J. consult. Psychol.*, 1959, **23**, 60–65.

LEWIN, K. The nature of field theory. In M. H. Marx (Ed.), *Psychological theory.* New York: Macmillan, 1951.

LITTLE, K. B., & SHNEIDMAN, E. S. Congruencies among interpretations of psychological test and anamnestic data. *Psychol. Monogr.*, 1959, 73(6, Whole No. 476).

LIVERANT, S. The use of Rotter's social learning theory in developing a personality inventory. *Psychol. Monogr.*, 1958, 72(2, Whole No. 455).

LOEVINGER, JANE. Theory and techniques of assessment. In P. R. Farnsworth (Ed.), *Annu. Rev. Psychol.*, 1959. Palo Alto, Calif.: Annual Reviews, 1959. Pp. 287–316.

LORD, E. E. Experimentally induced variations in Rorschach performance. *Psychol. Monogr.*, 1950, **64**(10, Whole No. 316).

MCCLELLAND, D. C., ATKINSON, J. W., CLARK, R. A., & LOWELL, E. L. *The achievement motive.* New York: Appleton-Century-Crofts, 1953.

MANDELL, ELIZABETH E. Construct validation of a psychometric measure of expectancy. Unpublished master's thesis, Univer. Colorado, 1959.

MANDLER, G., & SARASON, S. B. A study of anxiety and learning. *J. abnorm. soc. Psychol.*, 1952, **47**, 166–173.

MOSS, H. The generality of cautiousness as a defense behavior. Unpublished doctoral dissertation, Ohio State Univer., 1958.

MURRAY, H. A. Toward a classification of interaction. In T. Parsons & E. A. Shils (Eds.), *Toward a general theory of action.* Cambridge: Harvard Univer. Press, 1952.

MURSTEIN, B. I. A conceptual model of projective techniques applied to stimulus variations with thematic techniques. *J. consult. Psychol.*, 1959, **23**, 3–14.

MUSSEN, P. H., & NAYLOR, H. K. Relationship between overt and fantasy aggression. *J. abnorm. soc. Psychol.*, 1954, **49**, 235–239.

MUSSEN, P. H., & SCODEL, A. The effect of sexual stimulation under varying conditions on TAT sexual responsiveness. *J. consult. Psychol.*, 1955, **19**, 90.

NAYLOR, H. K. The relationship of dependency behavior to intellectual problem solving. Unpublished doctoral dissertation, Ohio State Univer., 1955.

NEFF, J. Individual differences in resistance to extinction as a function of generalized expectancy. Unpublished doctoral dissertation, Ohio State Univer., 1956.

PEAK, HELEN. Problems of objective observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences.* New York: Dryden, 1953.

PHARES, E. J. Expectancy changes in skill and chance situations. *J. abnorm. soc. Psychol.*, 1957, **54**, 339–342.

RORSCHACH, H. *Psychodiagnostics.* New York: Grune & Stratton, 1942.

ROTTER, J. B. Level of aspiration as a method of studying personality: II. Development and evaluation of a controlled method. *J. exp. Psychol.*, 1942, **31**, 410–422.

ROTTER, J. B. *Social learning and clinical psychology.* New York: Prentice Hall, 1954.

ROTTER, J. B. The role of the psychological situation in determining the direction of human behavior. In M. R. Jones (Ed.), *Nebraska symposium on motivation,* 1955. Lincoln: Univer. Nebraska Press.

SARASON, S. The test situation and the problem of prediction. *J. clin. Psychol.*, 1950, **6**, 387–392.

THOMAS, W. I. (Collected writings) In E. H. Volkart (Ed.), *Social behavior and personality: Contributions of W. I. Thomas to theory and social research.* New York: Social Science Research Council, 1951.

TOLMAN, E. C. *Purposive behavior in animals and men.* New York: Appleton-Century, 1932.

WEISSKOPF, E. A. A transcendence index as a proposed measure of projection in the Thematic Apperception Test. *J. Psychol.*, 1950, **29**, 379–390.

(Received February 2, 1959)

Psychological Review
1960, Vol. 67, No. 5, 317–331

LEARNING THEORY AND THE ACQUISITION OF VALUES[1]

WINFRED F. HILL

Northwestern University

The processes by which a child acquires the values of his culture and his various overlapping subcultures is, as a recent review of the subject points out (Dukes, 1955), still rather obscure. This obscurity is certainly not due to lack of interest in the topic. Psychologists, psychiatrists, anthropologists, sociologists, pediatricians, and educators have all given attention to the question of how children come to share the attitudes, ideals, and ethical standards of those around them. Nor is this interest undeserved, for few topics are of greater practical importance.

Perhaps this very convergence of interest from many directions is partly responsible for the difficulties involved in studying the topic. This area of research has become a battleground for conflicting terminologies, with one term often having a multiplicity of half-distinct meanings, and with what appears to be the same meaning often bearing different labels. Although many terms contribute to this confusion, three are of particular interest here: *identification, introjection,* and *internalization.* All involve some relation between an individual, hereinafter designated the subject (*S*), and another person or personalized entity, the model or *M*, such that *S*'s behavior is in some way patterned after *M*'s. However, these terms may refer either to a state of affairs or to the process which brought it about (Lazowick, 1955); the *M* may be a person, a group, or an idea (Glaser, 1958); and the relation may involve specific responses, broad meanings, or emotional reactions.

Some of the confusion as to the meaning of the term identification may be seen in the following uses. Lazowick (1955) distinguishes three main uses of the term identification in the literature: pseudoidentity, imitation, and personality change. He suggests that the term should be used only with regard to broad meanings, with imitation being the corresponding term for specific acts. Freud (1950) contrasts a boy's identification with his father, which forms the basis of his ego ideal, and his identification with his mother as an abandoned object cathexis. Davis and Havighurst (1947) maintain that a child will identify with his parents only if he loves them, but Anna Freud (1946) emphasizes identification with the aggressor. Lynn (1959) contrasts sex-role identification, which is "reserved to refer to the actual incorporation of the role of a given sex, and to the unconscious reactions characteristic of that role" (p. 127), with sex-role preference and sex-role adoption. He regards figure drawing as a measure of identification and choice of dolls for play as a measure of preference. Sears, Maccoby, and Levin (1957), however, use children's choices of dolls as a measure of identification. Sanford (1955), discouraged by such confusions of meaning, considers the possibility of abandoning the term identification altogether, but decides to retain it to describe a defense mechanism involving extreme adoption of *M*'s behavior by *S*, a mechanism which is not important in normal personality development. Finally, this collection of meanings, diverse as it is, still omits those cases where identification is used as a synonym for loyalty or for empathy.

---

[1] The research for this article was supported by a grant from the Carnegie Corporation to the Department of Psychology and the School of Education at Northwestern University. The author wishes to express his appreciation to Donald Campbell, Robert Winch, and the members of their seminar in social psychology for their many contributions to his thinking on this topic.

There is similar confusion concerning the meaning of introjection and internalization. Both carry the implication of values being incorporated into the personality. Hence, particularly with introjection, there is the suggestion of some relation with orality (Freud, 1950). However, Freud in the same discussion also uses introjection synonymously with identification. Parsons (1955), on the other hand, treats identification and internalization as synonyms.

The many discussions of these three terms in the literature seem to indicate that there are several processes involved but no generally accepted conventions for labeling them. A number of writers, including several already cited, have expressed discouragement at this state of affairs, but the usual result of such discouragement seems to be a redefinition of terms, which may clarify the particular exposition but which only serves further to confuse the field as a whole. Whereas Lynn (1959) believes that a term as widespread as identification must have potential usefulness, the present writer believes that clarity would be served by abolishing not only identification but also introjection and internalization from the technical vocabulary of personality development.

What, then, should be substituted? The topics to which the above terms have been applied certainly deserve discussion, and if the redefinition of the old terms is unsatisfactory, the introduction of new terms would be even worse. An answer may be found, however, in a sort of reductionism. Since the processes involved are learning processes, the existing vocabulary of learning is the obvious candidate for the job of describing them. It is quite possible, of course, that the existing vocabulary of learning theory will be inadequate for the complexities of value acquisition. However, if its use is carried as far as possible, the successes of this application should clarify our thinking about personality development, while any gaps which result should point to possible extensions of learning theory.

This approach involves treating human learning in a sociocultural environment in the same terms, at least for a first approximation, as animal learning in the environment of laboratory apparatus. For this purpose, the social rewards and punishments applied to humans may be treated as equivalent to the food pellets and electric shocks used with rats. Similarly, social roles are the equivalents of mazes which must be learned in order to obtain the rewards and avoid the punishments. Human beings of course constitute a far more variable environment than laboratory hardware, and one on which *S* can exercise greater influence. However, since most of the theory in this area is concerned with the adaptation of *S* to a relatively constant human environment (whether it be called culture, social system, or the personalities of the parents), this should not prove a serious stumbling block. There is ample precedent for such an approach in the writings of Dollard and Miller (1950), Mowrer (1950), Whiting and Child (1953), and others.

In addition, this approach treats values as nothing more than inferences from overt behavior. In principle this assumption should cause no difficulty. Few behavioral scientists would regard values (in the empirical, not the transcendental sense) as fundamentally different from such behavioristic constructs as Hull's (1943) habit strength or Tolman's (1949) equivalence beliefs. In practice, however, some theorists might take issue with this view on at least two bases.

For one thing, the measurement of values (including attitudes, ideals, and ethical standards) is commonly by verbal methods (see the review by Dukes, 1955). This leads to the suspicion that any measurement of values by nonverbal means must be inadequate, that only verbal measures can get at the significance of an act for an individual. However, verbal responses are part of the total behavior of the human organism and may be studied like other responses. The processes of unconscious motivation, semiconscious hypocrisy, and deliberate concealment all indicate that it would be unwise to treat verbal and nonverbal measures of values as equivalent. Rather than treat either verbal or nonverbal behavior as the true indicator of values and the other as a side issue, it seems more useful to study both and to ascertain empirically to what extent they lead to the same generalizations about a given *S*.

Another possible objection involves the distinction between specific acts and broad meanings, as in Lazowick's (1955) contrasting definitions of imitation and identification, noted above. This distinction between specific acts and broad meanings may refer either to the presence of mediating responses (see Osgood, 1953) or to the generality of the stimuli and responses involved. Neither of these distinctions, however, is dichotomous. Hull (1952) has indicated how a mediator, the fractional antedating goal response, may function in animal behavior, and Russell and Storms (1955) have demonstrated mediational processes in rote learning. Thus the mediation mechanism is by no means restricted to the "higher mental processes" of humans. As for the generality of the behavior and of the stimuli which guide it, this presumably represents a continuum from the most specific to the most inclusive categories. If, for example, washing the hands before meals is an example of imitation and cleanliness an example of identification, where would wearing clean clothes be classified? So, although the distinction between specific acts and broad meanings is a legitimate one, there are no sharp breaks on the continuum and there is no reason to assume that basically different laws are involved.

In view of the above considerations, an attempt to study the acquisition of values as a branch of learning theory appears justified.

KINDS OF REINFORCEMENT

The concept of reinforcement is basic to learning theory. While theorists are by no means unanimously agreed on the value of reinforcement terminology, there is little question that an empirical law of effect holds, that the consequences of an act influence its subsequent occurrence. A classification of kinds of reinforcers will be used here as the basis for analyzing the learning of values.

Primary Reinforcement

For the present purpose, three kinds of reinforcement may be distinguished: *"primary," secondary,* and *vicarious.* Placing "primary" in quotes indicates that it refers to the effects not only of innate physiological reinforcers but also of those social reinforcers which play a primary role in human motivation. Presumably the positively reinforcing effects of attention and praise and the negatively reinforcing effects of criticism, ridicule, and rejection are at least partly learned, but the nature of the learning process is obscure, and at the present level of analysis it seems preferable to treat praise for a human as comparable to food for a rat. The distinction between "primary" and secondary reinforcement is thus one of convenience between that which we take as given and that for which we can find a specific learned basis. Though arbitrary, this distinction is perhaps no more so than the decision as to whether food in the mouth should be considered a primary or a secondary reinforcer.

One particular kind of learning by "primary" reinforcement is the acquisition of a generalized tendency to imitate others. Miller and Dollard (1941) have indicated how a generalized tendency to imitate the behavior of others may be learned in the same way as any other class of responses. Although their demonstrations of imitation involved *S*'s patterning his behavior after a leader who was present, Church (1957) has shown that rats can also learn to respond appropriately to the same cues to which the leader rat is responding. In spite of some negative animal evidence (Solomon & Coles, 1954), there is little doubt that humans can learn to pattern their behavior after that of other people, not only when the *M* is present, but also in *M*'s absence by utilizing the appropriate environmental cues. As the child is repeatedly rewarded for imitative behavior in a variety of otherwise different situations, and as his capacity for abstraction increases, it seems plausible that a generalized imitative tendency would develop. It would be desirable, however, if the widespread anecdotal support for this deduction could be bolstered by experimental data.

The same process presumably applies to verbal instructions. The child is typically reinforced (though with some striking exceptions) for doing what others tell him to do. Hence (common parental impressions to the contrary notwithstanding), a generalized tendency toward conformity to verbal instructions may be expected to develop. With increasing intellectual development, this tendency should come to include conformity to fictional examples or to abstract ethical exhortations.

Secondary Reinforcement

Although no basic distinction is made here between primary and secondary reinforcement, there is one case frequently discussed in the literature where the acquisition of reinforcing properties by certain stimuli may be analyzed in detail. These stimuli are those which are connected with care of the child by adults, i.e., the nonessential aspects of nurturance. These include patterns of speech, facial expressions, gestures, and the like. Since these occur with those nurturant behaviors which are primary reinforcers, such as feeding and cuddling, they may become secondary reinforcers. By stimulus generalization, these behaviors should also be rewarding to the child (although less so) when produced by himself. As the child grows older and the parents expect him to take greater care of his own needs, he is more and more forced to provide not only his own primary reinforcers, but his own secondary reinforcers as well. Hence he may be expected to show some of the same mannerisms as his parents showed when caring for him.

This kind of learning appears to be one of those processes which Freud (1950) includes under the heading of identification or introjection, that in which abandoned object cathexes become incorporated into the ego. However, in the view presented here, the coincidence of abandonment and incorporation into the ego refers only to performance, not to learning. The secondary reinforcing value of the parental mannerisms is built up during the period of nurturance, but becomes evident in the child's behavior as nurturance begins to be withdrawn. This learning process has been discussed by Mowrer (1950, Ch. 24) in connection with the learning of language, and by Lindemann (1944) as a reaction to the death of a loved one. Although this process appears better adapted to the learning of rather trivial mannerisms, it is capable at least in principle of being adapted to more general and significant values as well.

Vicarious Reinforcement

Vicarious reinforcement does not have the same dignified status in learning theory as do primary and secondary reinforcement, but some such process appears necessary in order to explain some important human learning. Vicarious reinforcement involves the generalization of reinforcing effects from others to oneself, hence learning from the reinforcers which others receive. A given act is reinforced for *S* as a result of the act being performed by *M*, followed by reinforcement to *M*. For example, if *S* observes *M* trying to solve a problem by certain techniques and succeeding, *S* is more likely to use the same techniques when faced by a similar problem than if *M* had failed to solve the problem. Although most of the evidence for such learning is anecdotal, Lewis and Duncan (1958) have provided some evidence of it in a human gambling situation, and Darby and Riopelle (1959) have demonstrated it in discrimination learning by monkeys.

Although vicarious reinforcement involves selective imitation, it differs from the selectivity of imitation, described by Miller and Dollard (1941), in which *S* imitates some *M*s and not others because of differential reinforcement received by *S* for imitating the two *M*s. Vicarious reinforcement does not involve any reinforcers delivered directly to *S;* the discrimination of *M*s to be imitated or not is made entirely on the basis of *S*'s observation of *M*'s experience. This distinction is emphasized by Campbell (in press). In Hullian terms, vicarious reinforcement involves the acquisition of *K* by observation.

This type of learning need not be restricted to the effect of particular reinforcers administered to *M* under specific conditions. Stimulus generalization should occur not only from *M*'s behavior to *S*'s but also from one act of *M*'s to another. As a result, if *M* is frequently reinforced, *S* should find it rewarding to resemble *M* in general, including imitation of some of *M*'s behaviors which *S* has never seen rewarded. Thus a beginning salesman (*S*) might treat a customer with extreme politeness because he had observed another salesman (*M*) making a large sale while using such behavior, and he might also smoke a cigar because he had observed his highly successful salesmanager (*M*) doing so. In the former case *M*'s behavior (politeness) and *M*'s reinforcement (a large sale) were paired, whereas in the latter case the reinforcement (business success) was a perennial experience of *M*, but not paired with the particular behavior (smoking cigars) in question. Both, however, are examples of vicarious reinforcement.

Vicarious reinforcement corresponds to identification as defined by Masserman (1946) and to that aspect of identification referred to by Kagan (1958) as "the motivation to command or experience desired goal states of a model" (p. 298). Freud's (1946) identification with the aggressor also fits under this heading if successful aggression is assumed to be reinforcing to the aggressor. Sanford's (1955) concept of identification also involves the adoption as *M* of someone perceived by *S* as successful.

Conflicting Sources of Reinforcement

Traditionally the terms identification, introjection, and internalization might be applied to any or all of the learning processes described above or to their end product, similarity between *S* and some *M*. Since for the most part these processes involve learning by imitation, require some kind of reinforcement, and result in similarity between *S* and *M*, it may be asked why detailed analysis of the rather subtle differences among them is called for.

The answer is apparent when the possibility of conflict is considered. Conditions for a given *S* may be such that one of these processes tends to produce one kind of behavior while another tends to produce quite different or even opposite behavior. Such a conflictful situation might be expected in a child reared by a nurturant mother, whose mannerisms would become secondarily reinforcing, and a domineering father, who would be perceived as successful in mastering the environment. Freud (1950) recognized the frequent occurrence of just such a conflict, but did not consider it necessary to use different words for the two learning processes. Another common conflict is between the tendency to imitate *M*s whom *S* is directly reinforced for imitating (e.g., well behaved children) and the tendency to imitate *M*s whom *S* perceives as successful (e.g., tough kids). Such conflict is inevitable to a certain extent in children, since they are not permitted to imitate their (presumably more successful) elders in all respects, but it is particularly prominent in members of low-status social categories, (e.g., Negroes) who are often conspicuously not reinforced for imitating high-status *M*s. In the broadest sense, any situation where there is discrepancy among what *S* is told to do, what he is rewarded for doing, and what he sees others doing is a potential conflict situation, and one in which the use of any single inclusive term such as "identification" obscures the relevant variables.

The occurrence of conflict among the various reinforcement processes makes possible a finer analysis of the acquisition of values than could be made otherwise. If there is perfect agreement among what *S* is told to do, what those who nurture him do, what those around him conspicuously master the environment by doing, and what he himself is directly rewarded for doing, there is little basis for judging how much each of these factors contributed to *S*'s adoption of the values of those around him. By observing situations in which they conflict, greater knowledge of the efficacy of each kind of reinforcement may be obtained. Research in conflict situations might answer such theoretically and practically important questions as: "Does dominance or nurturance on *M*'s part do more to make *M* effective in modifying *S*'s values?"; "Do words and examples completely lose their efficacy if the appropriate behaviors, when elicited, are not reinforced?"; and "To what extent is behavior influenced by *M*s presented verbally (e.g., in literature)?"

There is of course no special merit to the classificatory scheme presented here. Except for the concept of vicarious reinforcement, the writer has avoided attaching distinctive labels to the learning processes described. The purpose of this discussion was to show how the terminology of learning theory can be applied to processes of value acquisition which have been described by personality theorists. This not only serves as a step toward the integration of these two areas of study, but also suggests the probable usefulness of employing such independent variables as number, percentage, magnitude and delay of reinforcement, distribution of practice, and discriminability of stimuli in the study of value acquisition. As both learning theory and personality theory develop further, it is to be expected that any schema developed now will be at least partially replaced by newer concepts. Rather than developing in further detail the ideas suggested above, the remainder of this discussion will therefore concentrate on the application of this kind of thinking to a narrower area, the development of conscience.

CONSCIENCE

Negative values, or conscience, have received much more attention than positive values. Educators seeking to improve children's characters, psychoanalysts concerned with the tyranny of the superego, anthropologists trying to distinguish between shame and guilt cultures, and experimental psychologists noting the persistence of avoidance responses have shared this emphasis on values of the "Thou shalt not" variety. Because of this widespread interest in conscience, it is a particularly appropriate topic with which to illustrate the possibilities of the learning theory approach to the study of values. Sears, Maccoby, and Levin (1957), in their challenging book *Patterns of Child Rearing*, devote a chapter to the development of conscience in preschool children. Their treatment of the topic will serve as a starting point for the present analysis.

Criteria for Conscience

Sears, Maccoby, and Levin give three criteria for recognizing the operation of conscience in young children: *resistance to temptation, self-instructions* to obey the rules, and *evidence of guilt* when transgression occurs. These three criteria are treated jointly as defining conscience, and no attempt is made to analyze their separate developments. Although the authors mention that the aspects of conscience do not necessarily all appear at once, they regard conscience as representing an internalization of control which is fundamentally different from external control, whether by force, fear of punishment, or hope of material reward. This treatment of conscience as essentially a single variable seems premature in our present state of knowledge; certainly the learning theory approach to personality advocated here would involve separate analyses of these diverse response patterns.

The first criterion, *resistance to temptation,* may be viewed simply as avoidance learning. Although this kind of learning is still a focus of theoretical controversy, much experimental data are available concerning it (Solomon & Brush, 1956). Sidman's (1953) studies of avoidance behavior without a warning signal and Dinsmoor's (1954) analysis of punishment show how feedback from an individual's own acts can become a cue for avoidance, and how persistent such avoidance may be. Although children can presumably learn to respond to more abstract characteristics of cues than can animals, there is no reason to regard a child's learning to avoid certain behaviors as fundamentally different from a rat's learning to do so. The fact that the child avoids the forbidden acts even in the absence of the parents is presumably due to the parents' having in the past discovered and punished (in the broadest sense of that word) transgressions committed in their absence.

This relating of conscience to avoidance learning suggests that independent variables known to be effective in animal avoidance learning would be among the most appropriate ones for study in connection with the development of conscience in children. Within certain limits, the greater the intensity of the punishments (Brush, 1957; Miller, 1951) and the shorter the delay between transgression and punishment, (Mowrer & Ullman, 1945; Solomon & Brush, 1956) the greater should be the resulting inhibition. Though the data are somewhat ambiguous, greater certainty of punishment might be expected to produce inhibition which would be more complete in the short run but also less persistent once punishment was permanently withdrawn (Grant, Schipper, & Ross, 1952; Jenkins & Stanley, 1950; Reynolds, 1958). This prediction suggests that even this one criterion of conscience may not be unitary, that different laws may apply depending on whether one asks how completely the child obeys the prohibitions or how long he continues to obey them after leaving the parental home. If partial reinforcement should turn out to be a crucial variable in the human situation, these two criteria might even be inversely related. The prediction also suggests that the question, "Is inconsistent discipline bad?" is far too simple; one must at least ask, "Bad for what?"

It must also be kept in mind that punishment is not restricted to physical chastisement or even to noxious stimuli in general, including scolding and ridicule. Withdrawal of positive reinforcers may be very effective as a punishment, a fact which complicates the analysis. As this is a much discussed topic in personality theory, it will be considered below.

Sears, Maccoby, and Levin's second criterion of conscience, *self-instruction,* obviously makes the human case different from the animal case, but it does not introduce any new motivational principle. One of the advantages of membership in the human species is the possibility of using verbal symbolization in dealing with one's problems. It is natural that a person learning an avoidance, like a person learning any other difficult response pattern, should give himself verbal instructions, especially since verbal coaching by others is so important in the learning of social prohibitions. Moreover, such self-instruction is an imitative act which might be learned according to any of the reinforcement paradigms discussed above. Presumably the learning of prohibitions proceeds differently in verbal and nonverbal organisms, but observations of the relation between moral statements and moral behavior (Hurlock, 1956, pp. 406, 411–412) argue against the assumption that there is a high correlation between verbal and other criteria of conscience, except as both are influenced by the values represented in the social environment.

The third criterion of conscience, *guilt* at violations of the prohibitions, is itself complex, with many verbal, autonomic, and gross behavioral aspects. However, the striking paradox about guilt, which has seemed to some students to set it apart from the ordinary laws of learning, is that it often involves the seeking of punishment. The person who has transgressed, rather than trying to avoid punishment, or even waiting passively for it to come, actively seeks out the authorities, confesses, and receives his punishment with apparent relief. He may also, or instead, go to great lengths to make restitution. Were it not for these phenomena of punishment-seeking and self-sacrificing restitution, it would be easy to dismiss guilt as merely the kind of fear associated with anticipation of certain sorts of punishment. As it is, the existence of guilt serves as an argument for regarding conscience as something more than the sum of all those avoidances which have moral significance in one's culture.

However, the attempt to distinguish between guilt-controlled and other behaviors has not been very successful. Though the distinction between guilt cultures and shame cultures has had a considerable vogue in anthropology (e.g., Benedict, 1946; Havighurst & Neugarten, 1955; Mead, 1950), the inadequacies of the distinction have been pointed out by Ausubel (1955) and by Singer (1953). Moreover, the relation between conformity to a standard and guilt when the standard has been violated is open to question. Shaw (1948) suggests that confession may even be so satisfying to some people that it constitutes a reinforcement for sinning. So, although the phenomena of guilt may raise difficulties for learning theory, these difficulties probably cannot be solved by using guilt to define a distinctive kind of learning.

The above considerations should suffice to indicate that conscience cannot be assumed a priori to be unitary. The extent to which short-run conformity, long-run conformity, self-instructions to conform, certain kinds of distress at having failed to conform, and voluntary confession of nonconformity are intercorrelated is a matter to be empirically determined. Moreover, even if high positive intercorrelations are found, it is possible that they may reflect correlations in the environment rather than any fundamental unity of process. If environmental pressures toward conformity vary markedly, artificially high correlations among the criteria of conscience are to be expected. However, even when this artifact is removed, an analysis of separate learning processes for different behaviors may still lead to the prediction of high correlations among the behaviors. Such an analysis is presented below.

Learning of Conscience

Sears, Maccoby, and Levin found that the development of conscience, as defined jointly by their three criteria, was greater in those children whose parents used love-oriented forms of discipline (praise, isolation, and withdrawal of love) than in those whose parents used "materialistic" forms of discipline (material rewards, deprivation of privileges, and physical punishment). A similar finding, though not highly reliable statistically, is reported by Whiting and Child (1953, Ch. 11) in a cross-cultural study of guilt as measured by attitudes toward illness. This is consistent with the widely held view that the acquisition of parental values occurs most fully in an atmosphere of love (e.g., Ausubel, 1955; Davis & Havighurst, 1947). It is possible, however, that this finding may be due, not to love-oriented discipline as such, but to other characteristics of discipline which are correlated with it. The effect of this kind of discipline may be to accentuate the learning of several different responses, all of which contribute to the overall diagnosis of high conscience.

The various kinds of punishments commonly applied to children probably differ markedly in the temporal relations and the reinforcement contingencies involved. Physical punishment is likely to occur all at once and be over quickly, while punishment by deprivation of objects or privileges is likely to be either for a fixed period of time or for as long as the disciplinarian finds convenient. Discipline by withdrawal of love, on the other hand, probably much more often lasts until the child makes some symbolic renunciation of his wrongdoing, as by apologizing, making restitution, or promising not to do it again. The child is deprived of his parents' love (or, the parents would claim, of the outward manifestations of it!) for as much or as little time as is necessary to get him to make such a symbolic renunciation. When he has made it, he is restored to his parents' favor. If the normal relation between the parents and child is one of warmth, such discipline strongly motivates the child to make the renunciation quickly. On repeated occasions of transgression, punishment by withdrawal of love, and symbolic renunciation, the child may be expected not only to learn the renunciation response as an escape from parental disfavor but eventually to use it as an avoidance rather than merely an escape response. Thus if the wrongdoing is not immediately discovered, the child may anticipate his parents' impending disfavor by confessing in advance and making the symbolic renunciation.

The result of this hypothesized sequence of events is that the child makes a verbal response which is in effect an instruction to himself not to repeat his wrongdoing. The next time temptation comes, he is more likely to make this verbal response before transgressing. Although this does not guarantee that he will not transgress, it is likely to reduce the probability. If he succumbs to temptation, he is more likely to confess before being caught and thereby avoid the temporary loss of his parents' love. Thus if the above reasoning is correct, all three criteria of conscience should be present to a greater degree in the child who has been disciplined in this fashion than in other children. According to the present hypothesis, however, this will be due to the fact that punishment continues until the child makes a symbolic renunciation, rather than to the fact that the punishment involves withdrawal of love. If physical chastisement or loss of privileges are used in the same way, the same outcome is predicted.

A possible weakness of this hypothesis is that children might learn a discrimination between the symbolic and the actual avoidances, so that they would develop a pattern of violating parental standards, immediately confessing and apologizing, and then transgressing again at the next hint of temptation. If forgiveness is offered freely and uncritically enough, such a pattern presumably does develop. In this case the correlation among the criteria of conscience would be expected to drop, actual avoidance of wrongdoing no longer being associated with the other criteria. (For this reason, Sears, Maccoby, and Levin might have found smaller relations if they had studied older children.) However, if the parents' discrimination keeps up with the child's so that the child cannot count on removing all the parents' disfavor with a perfunctory apology, the efficacy of this kind of discipline should be at least partially maintained.

If this explanation of greater conscience in children disciplined by withdrawal of love is correct, why was greater conscience also found with the other kinds of love-oriented control? Since these were all found to be intercorrelated, and since their relations to the degree of conscience were uniformly low, interpretations either of separate techniques or of love orientation as a general trait are necessarily somewhat dubious. As an example of the difficulties involved, it may be noted that reasoning with the child is counted as a love-oriented technique solely on the grounds of its correlation with the other such techniques. Nevertheless, it shows a higher relation to conscience than do two of the three clearly love-oriented techniques. In view of such complexities, it seems legitimate to suggest that the crucial factor in those techniques associated with conscience may not be love orientation as such, but something else correlated with it.

To test this hypothesis, it would be necessary to have further detailed information of the sort that Sears, Maccoby, and Levin used, so that disciplinary methods could be classified according to the time relations discussed above. It is predicted that the parents' tendency to make termination of punishment contingent on symbolic renunciation would be correlated with love-oriented discipline. However, if each were varied with the other held constant, conscience should be more closely related to response contingency than to love orientation.

Along with this overall analysis of conscience, more detailed analyses could be made of the various components of conscience. According to the present view, intercorrelations among these criteria would be moderate for the entire sample and low when method of discipline was held constant.

The learning sequence discussed above is only one of several possible explanations of the Sears, Maccoby, and Levin finding. By suggesting that the crucial causal factor is not the distinction between materialistic orientaton and love orientation, but another distinction correlated with it, the present hypothesis gains an advantage in objectivity and in practical applicability. Whether it also has the advantage of correctness must be empirically determined. The chief purpose of the present example is to point to the availability of such reductionist hypotheses in the study of values and to argue that they deserve priority in the schedule of scientific investigation.

Permanence of Conscience

It would be particularly desirable to have a follow-up study to compare evidences of conscience in kindergarten with those of the same people later in life, when they were no longer primarily under the direct influence of their parents. Such a follow-up would help to clarify the relation between short-run and long-run conformity discussed above. Is the child who thoroughly obeys all his parents' prohibitions also the one who sticks to these standards when his parents are no longer around and his new associates have different standards? Anecdotal evidence can be cited on both sides, though the bulk is probably in the affirmative. To the extent that current and later conformity are independent, what variables influence one more than the other?

Predictions from learning theory on this topic are by no means unambiguous. Nevertheless, two lines of reasoning may be suggested concerning the type of discipline likely to result in the most persistent avoidances. (Persistence here refers, not to absolute level of avoidance, but to relative lack of decrement in the strength of avoidance with time.)

The first line of reasoning is from the differences in the slopes of *generalization gradients* for different kinds of learning (Dollard & Miller, 1950). In most cases the contrast in slope is between approach (or excitatory) and avoidance (or inhibitory) tendencies. It appears, however, that the basic distinction is between response tendencies activated by innate and by learned (generally fear) drives (Miller & Murray, 1952). When stimulus conditions change, the resultant removal of cues for fear produces a greater weakening of response tendencies based on fear than of response tendencies based on other drives. Hence, the generalization gradient of responses and inhibitions based on fear is steeper than that of other responses and inhibitions. This implies that discipline based on fear should lose its efficacy more quickly than discipline based on rewards as distance from the disciplinarian or any other change in conditions increases. Since this difference in slope is found on continua both of distance (Miller, 1944) and of similarity (Miller & Kraeling, 1952; Murray & Miller, 1952), it seems reasonable to predict that it also applies to that complex continuum along which an individual makes the transition away from parental apron strings. It would follow from this analysis that of two inhibitions learned in childhood, equal in age and original strength, one learned from the threat of losing rewards would be more effective later in life than one learned from the fear-provoking threat of punishment.

In this analysis, the advantages of discipline by manipulation (including withdrawal) of reward would apply to any kind of reward, material or social, not merely to parental love. However, the desire to continue receiving love from the parents may persist after the child has outgrown the need for other parental rewards, such as gifts and privileges. Discipline by withdrawal of love, in an atmosphere of warmth, might therefore be even more effective than other forms of discipline by denial of reward in producing persistent avoidances.

The other line of reasoning, involving the *partial reinforcement effect,* argues for the persistence of conscience learned by the process outlined above, in which a symbolic renunciation of wrongdoing terminates punishment. Although the greater resistance to extinction of responses which have received less than 100% reinforcement has been demonstrated primarily with positive reinforcement, it applies to negative as well (Humphreys, 1939, 1940; Grant, Schipper, & Ross, 1952). Partial reinforcement is of course present with all kinds of discipline, since punishment depends on the parents' moods and on the social situation, as well as on the child's being caught. However, the above analysis of the kind of punishment which terminates when the child makes a symbolic renunciation of wrongdoing suggests that such discipline may involve an additional source of partial reinforcement. As was indicated above, the child may learn that he can avoid punishment by confessing and apologizing. When this happens, the avoidance starts to extinguish. However, the discerning parent then learns not to accept the apology, and the child is punished anyway. The child must then make a more vigorous and convincing symbolic renunciation than before in order to terminate the punishment. In addition, the discrimination he has made between the symbolic renunciation and the actual avoidance is broken down; punishment can only be prevented by actual avoidance of wrongdoing. If, however, after a period of obedience he once more transgresses and then confesses, he is likely again not to be punished. This starts the cycle of extinction and reconditioning of the avoidance response going again, thus continuing to provide a reinforcement schedule in which only part of the child's transgressions are punished.

To predict that such partial reinforcement will retard extinction is admittedly problematic, both because of the complexity of the avoidance paradigm and because the unpunished transgressions are assumed to occur in blocks rather than randomly. Nevertheless, the hypothesis deserves consideration, not only as a prediction from learning principles to personality, but also as a case where the needs of personality theory might guide research in learning.

Although these two lines of reasoning agree in predicting maximally persistent conformity to parental prohibitions by children reared in an atmosphere of parental warmth and disciplined by withdrawal of love, they differ in their other predictions. To test these various hypotheses separately would require both short-run and long-run analyses of the effects of a variety of parental discipline patterns. The following hypotheses might be tested: (*a*) that discipline by deprivations (whether of things, privileges, or love) has more persistent effects than discipline by noxious stimulation (whether physical or social); (*b*) that where the child is taught to confess and apologize for his transgressions, avoidance behavior will go through cycles of extinction and reconditioning; and (*c*) that punishing only part of a child's transgressions results in more persistent obedience than does punishing all of them.

SUMMARY AND CONCLUSIONS

It is suggested that the terms identification, introjection, and internalization be replaced by detailed analyses in learning-theory terms of the acquisition of values. A reinforcement framework for such analyses is outlined, and examples are presented dealing with the concept of conscience and the factors influencing its development. It is argued that this would simplify terminology, encourage more precise study, and further the integration of learning and personality theories.

This analysis, like all attempts to integrate the harder-headed and the softer-hearted portions of behavioral science, is open to attack from both sides. On the one hand it may be objected that the present treatment is too cavalier with the interpersonal and intrapsychic complexities of personality development, that the internalization of values and the identification of one person with another cannot be treated as though they were nothing but the simple learning of a rat in a maze. The answer to this objection is that no "nothing but-ism" is intended; it is an empirical matter both to determine how far the principles of learning (not necessarily simple) can go in explaining personality development and to decide how much the additional principles suggested by some writers actually contribute to our understanding of the phenomena in question. The attempt to catch too much complexity at a single stroke may retard rather than advance our understanding.

On the other hand it may be objected that the interpretations given here are untestable, that the variables involve such diverse and subtle behaviors over such long periods of time as to defy adequate measurement. Admittedly the questionnaire, interview, and brief-observation techniques used in this area leave much to be desired. However, as long as applied behavioral scientists are called upon to deal with questions of personality development, poor data to guide their decisions are better than none. Study of learning of values by humans, guided by the principles of learning based on both animal and human studies, has the potential to make vital contributions to many theoretical and applied areas of knowledge. It is hoped that the present discussion may contribute something to that goal.

REFERENCES

AUSUBEL, D. P. Relationships between shame and guilt in the socializing process. *Psychol. Rev.*, 1955, **62**, 378–390.

BENEDICT, RUTH. *The chrysanthemum and the sword.* Boston: Houghton Mifflin, 1946.

BRUSH, F. R. The effects of shock intensity on the acquisition and extinction of an avoidance response in dogs. *J. comp. physiol. Psychol.*, 1957, **50**, 547–552.

CAMPBELL, D. T. Social attitudes and other acquired behavioral dispositions. In S. Koch (Ed.), *Psychology: A study of a science.* Vol. 6. *Investigations of man as socius: Their place in psychology and the social sciences.* New York: McGraw-Hill, in press.

CHURCH, R. M. Transmission of learned behavior between rats. *J. abnorm. soc. Psychol.*, 1957, **54**, 163–165.

DARBY, C. L., & RIOPELLE, A. J. Observational learning in the rhesus monkey. *J. comp. physiol. Psychol.*, 1959, **52**, 94–98.

DAVIS, W. A., & HAVIGHURST, R. J. *Father of the man.* Boston: Houghton Mifflin, 1947.

DINSMOOR, J. A. Punishment: I. The avoidance hypothesis. *Psychol. Rev.*, 1954, **61**, 34–46.

DOLLARD, J., & MILLER, N. E. *Personality and psychotherapy.* New York: McGraw-Hill, 1950.

DUKES, W. F. The psychological study of values. *Psychol. Bull.*, 1955, **52**, 24–50.

FREUD, ANNA. *The ego and the mechanisms of defense.* New York: Int. Univer. Press, 1946.

FREUD, S. *The ego and the id.* London: Hogarth, 1950.

GLASER, D. Dynamics of ethnic identification. *Amer. sociol. Rev.*, 1958, **23**, 31–40.

GRANT, D. A., SCHIPPER, L. M., & ROSS, B. M. Effect of intertrial interval during acquisition on extinction of the conditioned eyelid response following partial reinforcement. *J. exp. Psychol.*, 1952, **44**, 203–210.

HAVIGHURST, R. J., & NEUGARTEN, BERNICE L. *American Indian and white children.* Chicago: Univer. Chicago Press, 1955.

HULL, C. L. *Principles of behavior.* New York: Appleton-Century, 1943.

HULL, C. L. *A behavior system.* New Haven: Yale Univer. Press, 1952.

HUMPHREYS, L. G. The effect of random alternation of reinforcement on the acquisition and extinction of conditioned eyelid reactions. *J. exp. Psychol.*, 1939, **25**, 141–158.

HUMPHREYS, L. G. Psychogalvanic responses following two conditions of reinforcement. *J. exp. Psychol.*, 1940, **27**, 71–75.

HURLOCK, ELIZABETH B. *Child development.* (3rd ed.) New York: McGraw-Hill, 1956.

JENKINS, W. O., & STANLEY, J. C. Partial reinforcement: A review and critique. *Psychol. Bull.*, 1950, **47**, 193–234.

KAGAN, J. The concept of identification. *Pyschol. Rev.*, 1958, **65**, 296–305.

LAZOWICK, L. On the nature of identification. *J. abnorm. soc. Psychol.*, 1955, **51**, 175–183.

LEWIS, D. J., & DUNCAN, C. P. Vicarious experience and partial reinforcement. *J. abnorm. soc. Psychol.*, 1958, **57**, 321–326.

LINDEMANN, E. Symptomatology and management of acute grief. *Amer. J. Psychiat.*, 1944, **101**, 141–148.

LYNN, D. B. A note on sex differences in the development of masculine and feminine identification. *Psychol. Rev.*, 1959, **66**, 126–135.

MASSERMAN, J. H. *Principles of dynamic psychiatry.* Philadelphia: Saunders, 1946.

MEAD, MARGARET. Some anthropological considerations concerning guilt. In M. L. Reymert (Ed.), *Feelings and emotions: The Mooseheart symposium.* New York: McGraw-Hill, 1950.

MILLER, N. E. Experimental studies of conflict. In J. McV. Hunt (Ed.), *Personality and the behavior disorders.* New York: Ronald, 1944. Pp. 431–465.

MILLER, N. E. Learnable drives and rewards. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951. Pp. 435–472.

MILLER, N. E., & DOLLARD, J. *Social learning and imitation.* New Haven: Yale Univer. Press, 1941.

MILLER, N. E., & KRAELING, DORIS. Displacement: Greater generalization of approach than avoidance in a generalized approach-avoidance conflict. *J. exp. Psychol.*, 1952, **43**, 217–221.

MILLER, N. E., & MURRAY, E. J. Displacement and conflict: Learnable drive as a basis for the steeper gradient of avoidance than of approach. *J. exp. Psychol.*, 1952, **43**, 227–231.

MOWRER, O. H. *Learning theory and personality dynamics.* New York: Ronald, 1950.

MOWRER, O. H., & ULLMAN, A. D. Time as a determinant in integrative learning. *Psychol. Rev.*, 1945, **52**, 61–90.

MURRAY, E. J., & MILLER, N. E. Displacement: Steeper gradient of generalization of avoidance than of approach with age of habit controlled. *J. exp. Psychol.*, 1952, **43**, 222–226.

OSGOOD, C. E. *Method and theory in experimental psychology.* New York: Oxford Univer. Press, 1953.

PARSONS, T. Family structure and the socialization of the child. In T. Parsons & R. F. Bales (Eds.), *Family, socialization and interaction process.* Glencoe: Free Press, 1955.

REYNOLDS, W. F. Acquisition and extinction of the conditioned eyelid response following partial and continuous reinforcement. *J. exp. Psychol.*, 1958, **55**, 335–341.

RUSSELL, W. A., & STORMS, L. H. Implicit verbal chaining in paired-associate learning. *J. exp. Psychol.*, 1955, **49**, 287–293.

SANFORD, N. The dynamics of identification. *Psychol. Rev.*, 1955, **62**, 106–118.

SEARS, R. R., MACCOBY, ELEANOR E., & LEVIN, H. *Patterns of child rearing.* Evanston, Ill.: Row, Peterson, 1957.

SHAW, G. B. Preface to "Androcles and the Lion." In *Nine Plays.* New York: Dodd, Mead, 1948.

SIDMAN, M. Two temporal parameters of the maintenance of avoidance behavior by the white rat. *J. comp. physiol. Psychol.*, 1953, **46**, 253–261.

SINGER, M. B. Shame cultures and guilt cultures. In G. Piers & M. B. Singer (Eds.) *Shame and guilt.* Springfield, Ill.: Charles C Thomas, 1953.

SOLOMON, R. L., & BRUSH, ELEANOR S. Experimentally derived conceptions of anxiety and aversion. In *Nebraska symposium on motivation IV.* Lincoln: Univer. Nebraska Press, 1956.

SOLOMON, R. L., & COLES, M. R. A case of failure of generalization of imitation across drives and across situations. *J. abnorm. soc. Psychol.*, 1954, **49**, 7–13.

TOLMAN, E. C. There is more than one kind of learning. *Psychol. Rev.*, 1949, **56**, 144–155.

WHITING, J. W. M., & CHILD, I. L. *Child training and personality.* New Haven: Yale Univer. Press, 1953.

(Received August 3, 1959)

Psychological Review
1960, Vol. 67, No. 5, 332–342

WORD VALUES, WORD FREQUENCY, AND VISUAL DURATION THRESHOLDS

RONALD C. JOHNSON, CALVIN W. THOMSON, AND GERALD FRINCKE
San Jose State College

A number of studies have shown a relation to exist between word frequency, word value, and visual duration thresholds. Certain issues have developed concerning the interpretation of the observed relations. Two types of interpretation of results can be distinguished:

1. Those interpretations that claim word frequency to be the major determinant of visual duration threshold. This point of view contends that the responses made to the tachistoscopic presentation of words are learned in the same manner that other responses are learned. Differences in the visual duration thresholds of words can then be accounted for in terms of word frequency so that the introduction of such tenuous and "unanchored" variables as perceptual selectivity and perceptual defense is a violation of the law of parsimony. Thus, the problems that exist with regard to the differential visual duration thresholds of words are problems in learning, not problems in perception.

2. Those interpretations that attempt to show that differences in the visual duration thresholds of words are due, all or in part, to differences in the affective qualities (values, goodness–badness, affective tone, emotional valence) attached to words. These affective qualities affect visual duration thresholds directly, through perceptual selectivity and/or perceptual defense. Visual duration thresholds of words are determined by variables that can be called perceptual variables instead of or along with those variables that usually bear upon learning.

We wish to center our introductory discussion around those few studies which, for us, best exemplify the two positions described above. Solomon and Howes (1951) take the position that differences between words can be accounted for on the basis of word frequency. The perceptual process does not differ in any fundamental way from the learning process. They say:

> Emotional factors undoubtedly operate to an important extent in the building of word frequencies in a given life history. In this way they would be related to word frequency and, indirectly, to the duration thresholds. . . . But to date we can find no evidence to suggest that emotional factors operate in the tachistoscopic situation independently of their effect on word frequency (p. 267).

According to Solomon and Howes (1951, p. 258) visual duration threshold can be accurately predicted from a knowledge of population-wide word frequency as given in the Thorndike-Lorge tables (1944). There is no need to bring in concepts such as perceptual selectivity or perceptual defense in order to explain differences in visual duration thresholds of words. To Solomon and Howes emotional factors operate only in producing idiosyncratic variation in word frequency. These differences in frequency account for differences in the visual duration thresholds of words such as those found between the visual duration threshold of a value-oriented word for subjects (*S*s) who score high as opposed to low on specific scales of the Allport-Vernon Study of Values. They do not have any other influence on visual duration threshold.

The second position concerning the relation between word frequency, word value and visual duration threshold is clearly expressed by Postman and Schneider (1951). Faced with the problem of salvaging value as a variable influencing visual duration threshold, they say that

> It may . . . be more profitable theoretically to regard both frequency of word usage and duration thresholds as dependent variables, both manifestations of more fundamental psychological properties attributed to the organism, such as "habits," hypotheses, or even, perhaps, "personal values" (p. 277).

While not denying that word frequency can account for a considerable proportion of the variance in visual duration thresholds of words, they claim that values (called directive factors in this paper) directly influence the visual duration thresholds of infrequent words.

Postman, in a later article (1953), suggests that the frequency of usage of a specific word is determined by its reinforcing qualities. He suggests that frequent as opposed to infrequent words differ systematically in affective tone. He presents evidence to show that the sheer number of pleasant words is much greater than that of unpleasant words. Postman says, "Social control over environmental stimuli will, then, tend to create a correlation between frequency and positive value" (p. 68). Whether called values, directive states, or affective qualities, certain emotive or affective aspects of words influence their general (e.g., Thorndike-Lorge) frequency. Affective quality determines frequency, not merely in an idiosyncratic way as Solomon and Howes (1951) suggest, but also in a very general sense for all words used by all individuals.

It has been hypothesized from this second position that values can influence visual duration threshold in two ways. The affective qualities of words influence visual duration threshold *directly* as in the case of an *S*'s Allport-Vernon score influencing his thresholds for infrequent value oriented words (Postman & Schneider, 1951). (It has been shown that the *S* who values an area such as aesthetics highly reports infrequent "aesthetic" words at lower thresholds than an individual who does not value aesthetics highly. However, as Solomon and Howes point out, this may be due to "idiosyncratic" differences in frequency of exposure.) The affective qualities of words influence visual duration threshold *indirectly,* since the frequency of general usage of any word is determined to a considerable extent by the affective tone of that word, with this frequency then acting as a determiner of visual duration threshold (Postman, 1953).

A third position might also be taken concerning the relations between word frequency, word value, and visual duration threshold. Postman (1953) touches upon this last possible interpretation. He says, "Finally, we cannot entirely discount the possibility that familiarity resulting from frequency may be in itself a source of positive value" (p. 68).

It appears, historically, as if the first point of view—that emphasizing word frequency, playing down the influence of value except as it produces idiosyncratic differences in word frequency—has won out. None of the three positions described above have been proven to be untenable, however. Indeed, we have all of the original questions left, plus a few new ones as well. The questions that we see as being testable and requiring further investigation are these:

1. Is word value related, in a general sense, to word frequency? All of the three positions described above would admit the existence of a relation between value and

frequency—they would diverge in opinion with regard to the generality of the relationship. It is true that a knowledge of word frequency allows us to predict visual duration threshold. But why is the frequent word frequent? Perhaps we should go one step beyond frequency to determine whether other attributes of words vary systematically with frequency.

2. If a general relation between value and frequency exists, does it exist because more positively valued words are used more frequently or because one increases positive affect merely by increasing frequency? Or does the dependent–independent relationship depend on the experimental procedure used?

3. If value and frequency are related in a general, not merely an idiosyncratic sense, then one would have to separate out the influences of each on visual duration threshold. Are there significant differences in visual duration threshold between "good" and "bad" words of equal population-wide frequency? Are there significant differences between frequent and infrequent words equal in affective tone?

These are the three problems that we have attempted to deal with in the series of experiments reported below.

Our first concern was to determine whether a general relation exists between word value and word frequency. As we noted above, there is reason to suspect that emotional factors do operate in the building of word frequency. In a large sample of words for which we have actual measures of goodness, the words in the semantic atlas prepared by Jenkins, Russell, and Suci (1958), one finds better than a two to one preponderance of good over bad words. While the words in this atlas are not a random selection of words in the English language, the sources from which the words were selected seem unlikely to be systematically biased in frequency of pleasant as opposed to unpleasant words. One is led to believe that this ratio is characteristic of the entire English language when examining proportions of pleasant to unpleasant words in a sample of 150 words selected at random from the Thorndike-Lorge tables for use in the experiments in Series I, discussed below. Of these 150 words, 95 were on the good end (4.00 or less) of the semantic differential, a ratio of approximately 2:1. It seems highly probable that the sheer number of pleasant words in the English language is far greater than the number of unpleasant words.

It seems equally likely that the frequency with which any specific word is used also depends to a considerable extent on its goodness or badness: hence on its reinforcing qualities. This question is amenable to determination. As Solomon and Howes (1951) state, "In the absence of further data it seems best merely to define word frequency for English words as the frequency of words in the Thorndike-Lorge tables" (pp. 264–265). If affective factors operate in the building of word frequency, then pleasantly toned words should have higher frequencies in the Thorndike-Lorge word count than unpleasantly toned words. Experiment I was designed to determine whether this was the case.

EXPERIMENT I

In the first of this series of experiments we attempted to determine whether the pleasantness of words—in this case measured by the good–bad scale of the semantic differential—was related to word frequency as measured by the Thorndike-Lorge tables.

We selected a sample of words from the Thorndike-Lorge tables. This sample of words consisted of two words selected at random from within every alphabetic category (except X) in the tables. These words were rated on the good–bad dimension of the semantic differential by 24 *S*s, all freshmen taking an introductory psychology course. The rank order correlation between the goodness and the L count frequency of these words was +.63. Another random sample of 50 words was taken from the Thorndike-Lorge word list in the manner described above. These words were rated by 28 *S*s, a different group than those who made the first set of ratings. The rank order correlation between goodness and L count frequency in this list was +.40, again significant to the .01 level of confidence. A third random sample of 50 words was drawn from the Thorndike-Lorge tables. These words were rated by 24 *S*s, none of whom had taken part in previous experiments. The correlation between L count frequency and goodness for this set of words was +.38. All of these three correlations are significant to the .01 level. The results of these experiments, regardless of differences in the magnitude of the correlations, indicate that there is a positive and significant correlation between word frequency and goodness. Value attributes of words are related to word frequency—and not merely in an idiosyncratic way, but in a far more general sense as well.

EXPERIMENT II

Solomon and Howes (1951) list 30 frequent and 30 infrequent words that they used in a study of the relation of word frequency to visual duration threshold. Five frequent and 5 infrequent words had to do with each of the six value areas of the Allport-Vernon Study of Values. Frequent and infrequent words were sometimes, but not always, synonymous. We paired these words, matching the first frequent with the first infrequent word, and so on, through the two lists. Position of presentation within pairs was determined by coin flip. These 30 pairs of words were presented to 34 freshman introductory psychology students, none of whom had taken part in any previous psychological experiments. *S*s were told to "encircle the most pleasantly toned word of each pair." In 26 of the word pairs, the more frequent word was chosen by the majority of the *S*s as the most pleasantly toned. (The 4 exceptions, in which the least frequent word was chosen as the more pleasantly toned, were the word pairs limousine-automobile, economics-assets, orchestra-ensemble, and celestial-heavenly.) A sign test (Guilford, 1956, p. 248) indicates that this preference for the more frequent word in each pair could occur by chance less than one time in a hundred. Again, the more frequent word is rated as the better word.

In the experiment discussed below we attempted to determine whether frequency and goodness were related when nonsense syllables were used as stimulus materials.

EXPERIMENT III

Twenty-two *S*s rated two separate visually represented lists of nonsense syllables for goodness, as measured by the semantic differential. An interval of one week separated the ratings of the two lists. Each list consisted of eight 100%, eight 47–53%, and eight 0% association value (Glaze, 1928) nonsense syllables in random order. Syllables were selected at random from within each block of ten Glaze syllables as listed in the *Handbook of Experimental Psychology* (Stevens, 1951, pp. 540–546) at each of the association values used. Then the eight syllables used in each list at each association level were drawn from this pool of randomly selected syllables. We believed that if frequency of exposure was related to goodness, then the higher the association value of the syllable, the better the syllable should be rated. This seemed likely, since the association value of nonsense syllables appears to be determined largely by the frequency of occurrence of the letter combinations in meaningful words (Underwood

1959). Table 1, showing two sets of syllables and their mean ratings of goodness, indicates this to be the case.

A Mann-Whitney (1947) test of the significance of differences between syllables of 100% and 47–53% association values in List 1 indicates that the 100% list is rated as somewhat better ($P < .10$). The difference in goodness between 47–53% and 0% syllables is significant ($P < .01$); Mann-Whitney tests of List 2 show the 100% and the 47–53% list to differ in goodness ($P < .01$); and the 47–53% list and the 0% list to differ significantly ($P < .05$) as well.

Just as the goodness and the frequency of meaningful words in the English language are related to one another, so also is goodness related to the association value of nonsense syllables; the higher the association value, the better the syllable.

Our data indicate that value and frequency are related. We have moved at least somewhat beyond the empirical fact that word frequency is related to visual duration threshold. We know something else about the stimulus qualities of frequent as opposed to infrequent words; frequent words are not only more frequent but are better as well. Our second question then becomes relevant. Does a manipulation of frequency cause systematic variation in value or do differences in value produce differences in word frequency, or does the independent–dependent status of each of these variables vary depending on the experimental procedures used? Experiment IV bears upon this question.

TABLE 1

Ratings of Nonsense Syllables of Differing Association Values on the Good–Bad Dimension of the Semantic Differential

100% Syllables	Rating	47–53% Syllables	Rating	0% Syllables	Rating
List 1					
BUL	4.72	WOB	4.59	ZEQ	4.91
FES	3.41	KAW	4.33	WUH	4.48
HON	2.11	DUX	4.33	XIW	5.36
JIN	2.98	BOZ	4.00	WUQ	5.26
RAC	3.98	VOK	3.46	QIH	4.87
ROV	3.74	GEY	2.93	GUV	4.39
SUR	2.72	TIZ	3.52	VEC	3.91
WIL	2.46	VOZ	4.30	XUR	5.28
List 2					
DUL	4.27	BEK	3.46	YUF	4.46
BAL	2.35	PIJ	5.17	QIF	4.44
LOV	1.44	WEP	4.71	ZUK	5.04
SAR	3.23	RIQ	4.46	ZIF	4.06
LIK	2.50	YAF	3.96	XAD	4.85
HAV	2.35	LUB	4.15	XUC	5.35
WAT	3.46	GOW	3.75	QUJ	5.00
ROV	3.54	RUV	3.83	GUQ	5.98

Experiment IV

Fourteen *S*s rated the goodness of twenty nonsense words. These words were from the list of words used by Solomon and Postman (1952) in their study of the effect of "built in" frequency on visual duration threshold. The words used were: JANDARA, AFWORBU, BIWOJNI, NANSOMA, ENANWAL, IKTITAK, SARICIK, ZABULON, CIVADRA, LOKANTA, KADIRGA, ADAFNAW, BORULCE, NIJARON, ENSHIMI, INDULAM, TAVHANE, UDIBNON, DILIKLI, and MECBURI. Mean goodness scores were obtained. One week later each *S* was tested individually. When the *S* entered the experimental room he was given the following instructions (from Solomon & Postman, 1952):

> This is an experiment concerning the effectiveness of repetition in learning to pronounce strange words correctly. It has a direct bearing on the problem of reading words in a foreign language, as compared to hearing the words spoken. In addition, we are interested in knowing whether the relative effectiveness of the two kinds of learning methods depends on general reading ability.
>
> We are going to give you a deck of cards On each card is printed a strange word. We would like you to look at each card carefully and then pronounce the word in the way it would be pronounced if it were a word in the

> English language. Proceed steadily from card to card, turning over each one after you have finished with it. Go right through the deck and then stop unless you have serious doubts about your pronunciations.

The experimenter (*E*) handed the *S* a pack of 90 3″ × 5″ cards. In this pack were the words listed above. The first five words above were presented 10 times apiece, the next five, 5 times apiece, the next five, 2 times apiece, and the last five, once each. The cards were shuffled thoroughly after each *S* completed this task, so that they were in a relatively random order.

All responses were taped in order to lend realism to the expressed purpose of this new situation. Since *S*s had participated in another experiment on the pronounceability of words, it seems likely that (although the *S*s had seen the words before while rating them on the semantic differential) they were misled as to the purpose of this experiment. Each *S* rated the nonsense words a second time following this pronounceability test.

Figure 1 indicates the mean preexposure and postexposture ratings for words of each frequency of exposure. While the differences between ratings prior to exposure were not significant, a Kruskal and Wallis *H* test (1952) shows that words of differing frequencies of exposure do differ significantly in the postexposure test ($P < .01$, $H = 7.83$).

In this experimental situation a manipulation of word frequency alters *S*s' evaluations of word goodness. One can, however, think of instances where the reverse might well occur. The most obvious example of this is in verbal conditioning, where the differential reinforcement of a class of words produces systematic variation in the frequency with which this class of words is emitted. We would suspect that in this case frequency of emission (and hence of reinforcement) alters the affective quality of the words. Reinforcement might then produce further variation of frequency.

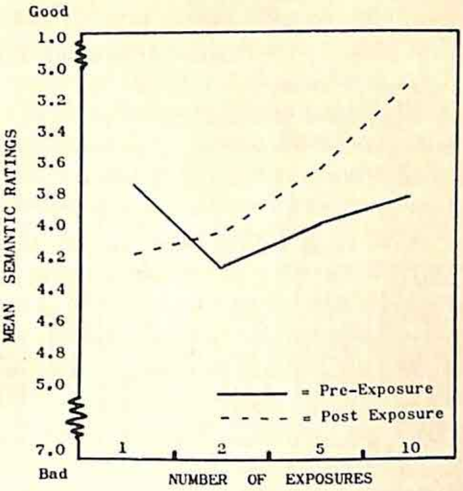


FIG. 1. Ratings of nonsense words before and after word frequency was manipulated.

No matter whether frequency of exposure alters value or vice versa, or both, one thing is evident—there is a relation between word frequency and *S*s' ratings of word goodness. Since this is so, our third question also becomes relevant. Since value and frequency are generally related, it became necessary to experiment with samples of words matched in frequency, varying in goodness, plus samples of words varying in frequency, matched in goodness, in order to separate out the influence of value and of frequency on visual duration threshold. This was attempted in Experiment V.

EXPERIMENT V

We had various groups of *S*s rate the goodness of a large number of words on the semantic differential. These groups rated 60 words of a Thorndike-Lorge G count of 12, 13, or 14, three random samples of 50 words each (described in Experiment I), and 39 words, one sample of 30 words plus

TABLE 2

WORDS USED IN THE TACHISTOSCOPIC EXPERIMENT

Words of Matched Goodness						Words of Matched Frequency					
Frequent			Infrequent			Good			Bad		
Word	*SD* Rating	Threshold	Word	*SD* Rating	Threshold	Word	G Count	Threshold	Word	G Count	Threshold
Sample 1											
Think	2.34	.0373	Income	2.52	.0500	Rally	12	.1092	Stammer	12	.1538
Butter	1.97	.0692	Caress	1.97	.1085	Glisten	13	.0777	Corrupt	12	.0985
Doctor	1.60	.0369	Ivory	1.68	.0262	Prosper	13	.1046	Defect	14	.0815
Health	1.97	.0492	Piano	1.97	.0546	Ivory	13	.0262	Hinder	13	.0969
Sleep	2.22	.0362	Radiant	2.08	.0662	Ginger	13	.0446	Wretch	12	.1285
Sweet	1.93	.0631	Patriot	1.90	.0992	Alert	14	.0392	Sneer	12	.0785
Mother	1.73	.0315	Vision	1.57	.0546	Deputy	13	.0446	Scandal	13	.1123
House	2.46	.0323	Agile	2.74	.0623	Vehicle	13	.0492	Dismal	13	.1162
Flower	1.87	.0369	Bible	1.87	.0462	Caress	12	.1085	Penalty	13	.0469
Sample 2											
Mountain	2.73	.0687	Bodice	2.74	.0965	Bodice	2	.0965	Slime	4	.0713
Water	2.43	.0217	Dough	2.47	.0757	Candy	32	.0409	Beggar	29	.1087
Window	2.68	.0400	Gradual	2.63	.0965	Engine	A	.0565	Bitter	A	.1704
Green	2.93	.0296	Elegant	3.00	.0852	Brave	A	.0739	Anger	A	.0417
Smooth	2.57	.0522	Zenith	2.62	.0330	Statue	32	.0748	Divorce	29	.0948
Father	2.36	.0800	Jelly	2.49	.0800	Cushion	21	.0900	Inferior	19	.1948
Pretty	2.13	.0661	Refresh	2.09	.0661	Garment	40	.1043	Grief	45	.0687
Street	2.77	.0496	Lenient	2.77	.1522	Church	AA	.0530	Danger	AA	.0583

9 words which happened to appear on the other lists, from Jenkins' semantic atlas. From these words, and from Jenkins' atlas, we selected 17 pairs of words matched in frequency, varying in goodness, and 17 pairs of words matched in goodness, varying in frequency. (Some words among those matched in goodness were rated by Jenkins' *S*s and by our own. Although the rank order correlation of goodness ratings between our own and Jenkins' sample of *S*s was +.98, for a group of 30 of these words, there was some variation in ratings between the two samples. Whenever this occurred, we used the mean goodness rating assigned by our *S*s.) The first 9 words from each list were presented to the first group of 26 *S*s; the last 8 to a second group of 23 *S*s. The four lists of words, divided as they were presented to two sample populations, are presented in Table 2.

The words were put on slides in the form of bold faced capital letters. The words were presented to *S*s by means of a Revere 888 tachistoscope. Each *S* sat 18 feet from the screen on which the words were projected. The testing room was fully illuminated while the tachistoscope was used at the smallest possible lens opening. The ratio of room illumination to screen illumination was 1:1.08. The words were presented in random order with the order remaining constant in all presentations to any specific *S* but with the order varying between *S*s. A modified method of limits was used in which each of the 36 words was presented once at each of the following speeds: 1/100, 1/50, 1/25, 1/5, 1/2, and 1 second. Under the conditions of presentation used in this experiment, only one word was recognized at 1/100 of a second while all words were recognized at 1 second.

Each *S*'s mean visual duration threshold for frequent as opposed to infrequent and good as opposed to bad words was obtained. The *S*s in the first sample ($N = 26$), exposed to the first group of words, differed significantly in mean thresholds for good vs.

bad words ($t = 3.56$, $P < .01$, using the formula for correlated data), with good words reported at lower thresholds. These *S*s also differed in mean thresholds for frequent vs. infrequent words ($t = 2.33$, $P < .05$), with frequent words reported at lower thresholds. The *S*s in the second sample ($N = 23$), exposed to the second group of words, differed significantly in mean thresholds for good vs. bad words ($t = 3.25$, $P < .01$), with good words reported at lower thresholds, and in threshold for frequent vs. infrequent words ($t = 3.82$, $P < .01$), with frequent words reported at lower thresholds.

DISCUSSION

We attempted to obtain data relevant to three specific questions. Let us look at these questions, along with the relevant data.

1. Is value related, in a general sense, to word frequency? Our data is unequivocally affirmative. Whether dealing with samples of words in the English language, with frequent vs. infrequent words in various Allport-Vernon value areas, or with nonsense syllables, frequency, as measured by the Thorndike-Lorge tables and by association value, and positive value, as measured by the good-bad scale of the semantic differential, are significantly related in a general sense.

2. If this general relation exists, does it exist because more positively valued words are used more frequently or because one increases positive value merely by increasing frequency? We found that a manipulation of frequency produces systematic variation in *S*s' ratings of the goodness of words. It seems highly probable, however, that the reverse may also be true—that by increasing the goodness of words, one also increases frequency. At this point it seems likely that the independent-dependent relation between goodness and frequency would depend on the experimental situation. We would not hazard a guess as to the effect of each on the other in general speech, although a relation between frequency and goodness has been shown to exist.

The data gathered in the first three experiments suggest that other attributes of the stimulus vary systematically as one varies frequency. Perhaps this is true for other forms of learning also. The fourth experiment would suggest the possibility that a manipulation of the frequency of occurrence of a stimulus and a contiguous response, in itself and without any regard to drive reduction, changes the reinforcing qualities of the stimulus. Perhaps we should look beyond the influence of frequency on response probability and ask *why* frequency changes response probability. Is it the difference in association value (and hence, in frequency of occurrence) that causes 100% association value syllables to be learned more rapidly than 0% syllables? Or is it the variation in affective quality accompanying differing frequencies of occurrence? Since the relation between goodness and frequency is not perfect, one can obtain syllables that vary from 0 to 100% in association value, yet are equal in semantic rating—this should provide a test for our question.

These results seem to have considerable social consequence. If word frequency and word goodness are related, and if a manipulation of the frequency of occurrence of a word produces systematic variation in word value, then one wonders whether ideas and personal values are manipulable in the same fashion. If so, then perhaps those elements of society that believe in censorship would find support in these experiments for one of their major assumptions, viz., that frequent exposure to an idea or a set of behaviors—violence, drinking alcoholic beverages, or what have you—lessens the "evil" associated with the idea of behavior. The validity of their second major assumption—that if we make a thing evil enough, people won't do this thing—is, of course, a very different problem.

If the more frequent becomes demonstrably better merely because it is more frequent, one can well understand the values of advertising. If it is generally true that the most frequent words in the English language are also the most good, the logical conclusion that this would lead to is that ads would be more effective if written at an even lower vocabulary level than is currently used. "Oh, brave new world. . . ."

3. If value and frequency are related, what are the influences of each on visual duration threshold when the other variable is held constant? Words equal in frequency, varying in value, are correctly reported by *S*s at different mean thresholds, with good words being correctly identified at significantly lower thresholds than are bad words. Words equal in goodness, varying in frequency, are reported by *S*s at different mean thresholds, with frequent words being correctly identified at significantly lower thresholds.

While it has been generally accepted that values have some secondary effect on visual duration thresholds, it is also generally accepted that the influence of values on threshold is small, producing idiosyncratic variations in frequency and hence influencing thresholds in this way. Our results indicate that values influence threshold even when frequency of exposure is held constant, and that this influence is approximately equal to that of frequency with value held constant, using these lists of words.

Since word goodness, independent from frequency, does significantly influence visual duration threshold, we are left with the problem of "Why?" The terms "perceptual selectivity" and "perceptual defense" are descriptive, not explanatory. We have no answers here, although we would like to discuss several possible explanations.

One interpretation of the fact that good words are perceived at lower threshold than bad words of the same frequency might be called the "simple" guess theory. Good words appear to be about twice as common as bad words in the English language. Specific good words are likely to be more frequently used than specific bad words. It seems, then, that good words occur between three and four times as frequently as bad words in terms of usage. Even though the frequency of occurrence of the stimuli are equaled, by matching good and bad words for Thorndike-Lorge frequency, response probabilities probably are still not equal. If an *S* sees a word in a tachistoscope but identifies only a "g—— d——" he is probably three or four times more likely, on the basis of responses made in the past, to fill in the gaps with a good word, such as "gradual," rather than a bad word, such as "goudge." This simple guess explanation may be partially correct but is not too satisfying to us, since an examination of our *S*s' responses to words on the good–bad lists indicates that there seems to be a rather strong likelihood that when *S*s guess at but miss a bad word, they respond with another bad word—incorrect but still bad, while when they respond incorrectly to the tachistoscope presentation of a good word, the incorrect response is usually another good word. This finding runs contrary to the idea that perceptual defense is a significant variable but also suggests that something more systematic than mere guessing is involved in errors and in correct responses to tachistoscopically presented words.

A second explanation for the differences in *S*s' mean visual duration thresholds might be called the "complex" guess theory. This explanation is derived from Bruner and Postman's (1949) hypothesis theory of perception and cognition. According to Postman (1951) hypotheses are, "in the most general sense, expectancies or predispositions of the organism which serve to select, organize, and transform the stimulus information that comes from the environment" (p. 249). The strength or availability of certain hypotheses is stronger than that of others. This strength is a function of the following factors: frequency of past confirmation of the hypothesis, number of alternative hypotheses, motivational relevance, and cognitive support. The first two factors—frequency of past confirmation and number of alternative hypotheses—fit well into the simple guess theory, while the last factor, that of cognitive support, may not be relevant to this discussion. The factor of motivational relevance might possibly, however, produce differential predictions between the two explanatory devices that we have labeled simple and complex guessing. As Postman (1951) puts it: "The stronger the motivational support for a given hypothesis, the less of appropriate information is required to confirm it" (p. 255). We would limit this statement somewhat by tying it only to the effect of reward or reinforcement on the number of cues required for the confirmation of a hypothesis. It may be that the differential reinforcement of various stimuli causes those stimuli (as well as those responses) that have most frequently led to reinforcement in the past to be more readily observable or available in the present. The availability of a specific response may well depend on the frequency that this response has been rewarded in the past—but perhaps because that aspect of the stimulus situation that has called forth the rewarded response in the past has become more salient—has become more of a "figure" as opposed to the other stimuli becoming more like the "ground." To paraphrase our quotation from Postman given above: The more reinforcing or rewarding a word is, the less cues are needed for the recognition of this word. We are advancing the proposition that reinforcement not only influences the probability that a given response will be emitted, but also that reinforcement causes that aspect of the stimulus situation that is salient at the time that the rewarded response is made to become more likely to be salient at the next presentation of the stimulus. If this is true, then even if word frequencies of good and bad words are equal and even if habitual modes of response (such as guessing good words) are controlled, the stimulus qualities of rewarding or reinforcing words would be such that less information is needed before the word is correctly recognized than would be the case for words which were presented with equal frequency but were not equally reinforcing. Something that might be called perceptual sensitivity might then be said to exist. This sensitivity would not, however, be a unique perceptual process, but would instead depend on those old standbys in all forms of learning: frequency and reinforcement, a little intermingled to be sure, but still observably present.

We have no way of testing out the "why" question with these data, although the types of wrong guesses made in our study incline us away from the simple guess theory. We believe that the data presented in this study are valuable chiefly in reviving the concept of perceptual selectivity, not in solving the problem of why what is called perceptual selectivity exists. By systematically varying frequency of exposure and frequency of reinforcement and nonreinforcement of nonsense syllables in future experiments, we hope to eventually be able to deal with the question of why good words are perceived at lower mean thresholds than are bad words.

SUMMARY

Three positions may be taken concerning the relation between word value and word frequency, and the relation of each of these to visual duration thresh-

old. Two of these points of view would suggest a general relation to exist between word frequency and word value. Three experiments demonstrate the existence of this relationship. The more frequent a word or a nonsense syllable occurs in the English language, the better it is likely to be rated on the good–bad scale of the semantic differential. Since this relation does exist, we attempted to test out, in Experiment IV, one possible explanation of *why* the relation exists. A manipulation of the frequency of nonsense words produces systematic variation in the rated goodness of these nonsense words. Words that are frequent are also very likely to be rated as good. Hence, to separate out the influence of word value and of word frequency on visual duration threshold, we tachistoscopically presented lists of words matched in frequency, varying in goodness; matched in goodness, varying in frequency. Subjects reported the good words at significantly lower thresholds than the matched bad words; the frequent words at significantly lower thresholds than matched infrequent words. Both frequency and value appear to operate in producing differential visual duration thresholds for words. Possible explanations and certain implications of these findings were discussed.

REFERENCES

BRUNER, J. S., & POSTMAN, L. Perception, cognition, and behavior. *J. Pers.*, 1949, **18**, 12–31.

GLAZE, J. A. The association value of nonsense syllables. *J. genet. Psychol.*, 1928, **35**, 255–267.

GUILFORD, J. P. *Fundamental statistics in psychology and education.* (3rd ed.) New York: McGraw-Hill, 1956.

JENKINS, J. J., RUSSELL, W. A., & SUCI, G. J. An atlas of semantic profiles of 360 words. *Amer. J. Psychol.*, 1958, **71**, 688–699.

KRUSKAL, W. H., & WALLIS, W. A. Use of ranks in one criterion variance analysis. *J. Amer. Statist. Ass.*, 1952, **47**, 583–621.

MANN, H. B., & WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. math. Statist.*, 1947, **18**, 50–60.

POSTMAN, L. Towards a general theory of cognition. In J. H. Rohrer & M. Sherif (Eds.), *Social psychology at the crossroads.* New York: Harper, 1951. Pp. 242–272.

POSTMAN, L. The experimental analyses of motivational factors in perception. In J. S. Brown (Ed.), *Current theory and research in motivation.* Lincoln: Univer. Nebraska Press, 1953. Pp. 59–108.

POSTMAN, L., & SCHNEIDER, B. H. Personal values, visual recognition, and recall. *Psychol. Rev.*, 1951, **58**, 271–284.

SOLOMON, R. L., & HOWES, D. H. Word frequency, word value, and visual duration thresholds. *Psychol. Rev.*, 1951, **58**, 256–270.

SOLOMON, R. L., & POSTMAN, L. Usage as a determinant of visual duration thresholds of words. *J. exp. Psychol.*, 1952, **43**, 195–201.

STEVENS, S. S. (Ed.) *Handbook of experimental psychology.* New York: Wiley, 1951.

THORNDIKE, E. L., & LORGE, I. *The teacher's word book of 30,000 words.* New York: Bureau of Publications, Teachers College, 1944.

UNDERWOOD, B. J. Verbal learning and the educative process. *Harvard educ. Rev.*, 1959, **29**, 107–117.

(Received September 15, 1959)

VOL. 67, NO. 6 NOVEMBER 1960

THE PSYCHOLOGICAL REVIEW

RATING SCALES, DISCRIMINABILITY, AND INFORMATION TRANSMISSION[1]

W. R. GARNER

Johns Hopkins University

Rating scales have been so widely used in psychology, both as research tools and in practical applications, that the determination of the optimum number of rating categories has been an important research problem in itself. The usual attack on this problem has been to determine the effect of number of rating categories on interrater reliability, and research on these lines has been well summarized by Guilford (1954, p. 289 ff.).

The purpose of a rating scale is to allow a rater or group of raters to demonstrate perceptual discrimination of a set of stimuli or objects to be rated. In other words, our concern is not with the rating scale per se, but rather with the extent to which different rating scales allow us to measure discrimination between the stimuli rated. And the basic question about number of rating categories is simply whether for any given rating problem there is an optimum number of rating categories, or at least a number of rating categories beyond which there is no further improvement in discrimination of the rated objects.

Techniques directly concerned with discriminability of stimuli have rarely been used for this problem, and yet since the problem basically is one of stimulus discrimination, it seems most appropriate that they should be.

Two techniques which have been successful in dealing with related problems are the use of information transmission measures and the use of discriminability scaling procedures. Garner and Hake (1951) described how information transmission measures can be used with the absolute judgment problem to determine the effective discrimination between the stimuli of a set. While there are ramifications of the absolute judgment procedure which are different from those of the rating scale situation, the problems are sufficiently alike that we can expect information transmission measures to be useful. Bendig and Hughes (1953) and Bendig (1954) have in fact used the information paradigm for the rating scale problem, and came to the conclusion that discrimination increases with an increase in number of rating categories up to at least 11 categories.

---

[1] This research was done in part under Contract N5-ori-166, Task Order 1, and in part under Contract Nonr-248(55), between the Office of Naval Research and Johns Hopkins University. This is Report No. 6 under the latter contract. Reproduction in whole or in part is permitted for any purpose of the United States Government. Appreciation is due to Robert Austin, Allen Madans, and Natalia Potanin for aid in various calculations.

In the same paper, Garner and Hake (1951) described a discriminability scaling procedure and pointed out the relatedness of this procedure to the use of information transmission measures. To our knowledge, however, discriminability scaling procedures have not been used in determining the optimum number of rating categories.

The purpose of the present paper is to show how both information measures and discriminability scaling procedures can be used to aid in the solution of the rating scale problem. In addition, it will be shown that these two procedures are quite closely related, and that measures from one can be used to estimate the other. An illustrative set of data will help clarify the nature of these relations.

The Data

The data used in these experiments were ratings of the legibility of handwriting samples. We obtained 20 different samples of handwriting of "The quick brown fox jumped over the lazy dogs," chosen to give us a reasonable spread of legibilities. The samples were arranged on a single mimeographed sheet, one to a line. A total of 50 raters then rated each sample on a rating scale with 4, 6, 8, 10, 12, 16, and 20 categories. Each rater rated all samples with each rating scale, using one rating scale until all samples were rated on that scale. Different raters used the rating scales in different counterbalanced orders, and the samples of handwriting were randomized differently on each sheet to minimize any time errors.

The raters were randomly selected from available graduate and undergraduate students. Since each of 50 raters rated each of 20 samples of handwriting on each of seven different rating scales, we had a total of 7,000 ratings as our basic data.

Information Transmission

The first analysis was to determine the information transmission, $U(R{:}S)$, between stimuli, or samples, and responses for each of the rating scales. (While we are using the term information transmission, we will use the more general notation of Garner and McGill, 1956, in which $U(R{:}S)$ is a contingent uncertainty between responses and stimuli.) All 50 ratings for each sample were pooled for this analysis. Since $U(R{:}S)$ is limited by the number of response categories, the various rating scales allowed maximum information transmissions from 2.00 bits for four categories to 4.32 bits for 20 categories.

The open squares of Figure 1 are the information transmission measures computed from the basic data, and these data show a consistent rise in information transmission as the number of rating categories is increased. This result is in agreement with the results of Bendig and Hughes (1953) and Bendig (1954). These calculations, however, are subject to a bias in the information measure which changes progressively with number of rating categories.

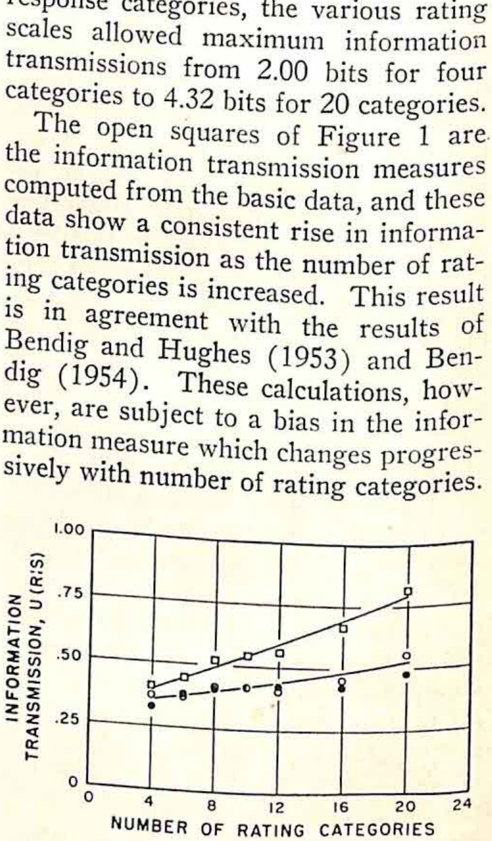


Fig. 1. Relation between number of rating categories and information transmission. (The open squares represent the direct measures of information transmission. The open circles represent the same data corrected for bias by the Miller-Madow correction. The two lines are least-squares fits for these two sets of measures. The closed circles represent estimates of information transmission obtained from discriminability functions.)

Each matrix has the same total number of ratings, but when the number of rating categories is increased, these ratings are distributed more thinly over the larger number of cells in the matrix. Miller (1955) has described a correction for this bias which is known as the Miller-Madow correction, and had himself suggested (Quastler, 1955, fn., p. 16) that Bendig's results would have been different if the correction for bias had been applied.

The correction for bias of information transmission is always in the direction of decreasing the estimate of transmission, and the fewer the number of cases per cell, the greater the correction; i.e., the greater does the directly calculated information transmission exceed the true value. When the Miller-Madow correction is applied, the result is that shown by the open circles of Figure 1. This result, which presents more nearly the true picture, shows considerably less effect of number of rating categories than does the result using raw measures. The correction flattens out the function. However, it is clear that there still is some effect of number of rating categories, and that an increase in the number of categories increases information transmission even up to 20 categories.

The two lines drawn in Figure 1 are linear functions, but they should be considered as entirely empirical description, since logically the functions cannot be linear much beyond the range of conditions investigated. For example, if only one rating category is used, there can be no information transmission, but the functions drawn do not extrapolate to zero with one rating category. Somewhere below four categories they must slope rapidly downwards.

On the other hand, there cannot be a meaningful increase in information transmission if more than 20 rating categories are used. With 20 or more rating categories, each rater can put each stimulus in a different category, and the rater can evidence no more discrimination than to rate each stimulus differently. And clearly the degree of discrimination by the group is limited by the discrimination of each rater. Actually the directly computed information transmission can increase with an increase in number of rating categories until each rater-sample combination produces a different rating. The function corrected for bias, however, would not increase concomitantly, since the bias correction would increase at approximately the same rate as the increase in number of rating categories. In the limit, the information transmission corrected for bias becomes degenerate, since with a large enough number of categories and a small enough number of ratings, it is impossible to demonstrate that the measured information transmission is different from the value expected on the assumption of no true contingency.

DISCRIMINABILITY FUNCTIONS

The second method of determining the amount of discrimination in the ratings is to construct discriminability scales for the set of stimuli. The dispersion of the scale values is then a measure of the relative discrimination between the various stimuli. We have carried out the necessary scaling procedures for these data in two different ways.

Pooled Data

With the first method, the data were kept in the pooled form in which they were used to compute information transmission. The S-R matrix of frequencies was converted into a set of equivalent pair comparisons with the technique described by Guilford (1938). The essence of this procedure is that the distribution of responses for each stimulus is compared category by category with the distributions for each of the other stimuli to obtain a measure of the equivalent number of times that each stimulus would have been rated greater than each other stimulus, assuming the ratings to be uncorrelated. From the matrix of equivalent pair comparisons, a discriminability scale was then constructed following the usual procedure for pair comparisons. One such discriminability scale was constructed for each rating scale.

Individual Data

With the second method, the ratings were changed to equivalent pair comparisons one rater at a time. Thus each rater's ratings of the 20 stimuli on one rating scale were changed into equivalent pair judgments in much the same way as Guilford (1954, p. 183 ff.) describes for converting ranks. Ties were used where necessary. These individual pair comparisons were then pooled for all 50 raters, and a discriminability scale constructed—again one for each rating scale. We thus have 14 discriminability scales, one for each of seven rating scales with each method.

The purpose in carrying out the two kinds of analysis was to determine whether the pooling of the data for the informational analysis had had any systematic effect on the function relating information transmission to number of rating categories. With a complete multivariate analysis of data such as these, we can determine separately the extent to which responses can be predicted from samples, from raters, and from the interaction between samples and raters. With only one rating per sample per rater, however, we have no choice but to pool the data, and are in effect accepting the interaction term as an error term rather than as a predictability term. We are in effect also accepting the contingent uncertainty between responses and raters as an error term. This last factor, which indicates the extent to which different raters use different responses *on the average,* could change with an increase in number of rating categories. For example, if all raters used rankings when 20 categories were allowed, then there could be no difference in use of responses by different raters, but such differences could exist with smaller members of categories. Thus, although it is not feasible to measure such a term directly, it is possible that the effect of this factor systematically decreased with the increase in number of rating categories, with the resultant increase in the information transmission which was obtained.

The first discriminability scale analysis used the pooled ratings to convert to equivalent pair comparisons. Thus any systematic differences between raters in ratings would still affect the final scale values. With the second scaling procedure, any such differences were removed by changing the data to equivalent pair comparisons before pooling.

Correlations of the Scales

The various scales produced by the two different methods are essentially identical. When the scale values are correlated between the two data processing methods for each of the seven rating scales, the seven correlation coefficients are all above .99. In addition, the correlations between scales for the different rating methods are quite high, as shown in Table 1. We have shown in this table only the correlations for the pooled data. Since all of the correlations between the two scaling methods are so high, a table of correlations using the individual data is practically identical to Table 1.

While the correlations between dis-

TABLE 1

CORRELATIONS BETWEEN DISCRIMINABILITY SCALE VALUES FOR DIFFERENT NUMBERS OF RATING CATEGORIES

	Number of Rating Categories					
	4	6	8	10	12	16
6	.90					
8	.90	.91				
10	.93	.96	.95			
12	.93	.97	.95	.96		
16	.90	.96	.92	.97	.97	
20	.89	.94	.90	.95	.96	.97

Note.—Pooled data were used for these correlations.

criminability scale values obtained from different numbers of rating categories are all quite high, it is clear that rating scales similiar in number of categories correlate more highly than do widely different scales. Thus the 4 scale and the 20 scale have the smallest correlation, while the 16 scale and the 20 scale have the highest correlation, along with some other close pairs.

Discriminability

These data on correlations between discriminability scale values show only that the various scales produced are highly predictable, one from the other. They do not show that the total amount of discrimination is the same. In order to determine the total amount of discrimination, the simplest procedure is to compute the standard deviations of the scale values for each of the scales. The scale whose stimulus values are most spread apart is the scale with the greatest interstimulus discrimination. Consequently, standard deviations have been computed for each of the 14 discriminability scales, and these values are plotted in Figure 2.

These data, like the information transmission data, show an increase in discrimination with an increased number of categories on the rating scale. And while the individual data show slightly greater discrimination than do the pooled data, the difference is very small—mean standard deviation for the pooled data is 0.881 and for the individual data it is 0.897.

What is more important, however, is that there is no evidence that the increase in discrimination with the increase in number of rating categories is in any way due to the pooling of the data. Both sets of points show a definite increase, and only for the 4 category scale is there any appreciable difference between the two standard deviations.

Once again the line drawn is a least-squares solution which can be considered only an empirical description. If one rating category is used, the standard deviation must be zero, and the curve shown does not extrapolate to zero at that point. Likewise, with a larger number of rating categories the scale values could not increase substantially, since no greater precision in equivalent pair comparisons could be obtained if there are more rating categories than there are stimuli. In other words, since the discriminability scale values are obtained from equivalent pair comparisons, the larger range of

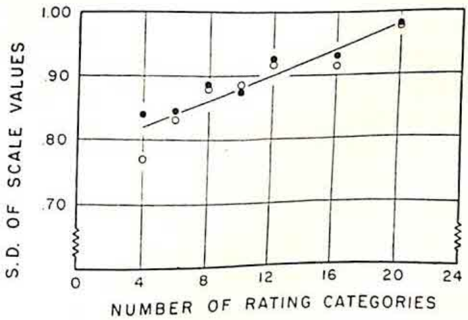


FIG. 2. Relation between number of rating categories and standard deviation of the discriminability scale values. (The open circles are for the data processed as a group; the closed circles for the data tabulated by individuals. The line is a least-squares fit.)

numbers in no way affects the standard deviations of the scale values.

ESTIMATES OF INFORMATION TRANSMISSION

We have shown that two alternative methods of determining the amount of discrimination in a set of stimuli evidenced by a rating procedure lead to the same conclusion, namely, that discrimination increases with an increase in number of rating categories, up to 20 categories in our case. It seems very reasonable that these two measures should, therefore, be related. It is the purpose of this section to show that they are related and that one can be quite exactly estimated from the other.

The actual equation for estimating transmitted information from the standard deviation or variance of the discriminability scale values is

$$\text{est}U(R{:}S) = \tfrac{1}{2}\log_2 [V(\bar{S}) + 1] \quad [1]$$

where $\text{est}U(R{:}S)$ is the estimated information transmission and $V(\bar{S})$ is the variance of the discriminability scale values of the stimuli. The $\bar{S}$ notation emphasizes the fact that these are mean scale values.

For the pooled data shown in Figure 2, an estimate of the information transmission has been computed for each number of rating categories, and these values are shown as the closed circles of Figure 1. It is obvious that Equation 1 provides a very accurate estimate of the unbiased information transmission, and that the two measures are highly related.

The rationale of Equation 1 can be understood from two related points of view. First, we can conceptualize the discriminability scale values as a set of values for a set of stimuli, and this same set of values as the mean numerical responses which would be obtained if both the stimuli and the responses were metricized in the unit normal system used with discriminability scales. In other words, we really have an *S-R* matrix, but one in which these new scale values, rather than the respective original designations, are used for both stimuli and responses. Now with such a matrix both the stimulus scale and the response scale are continuous, but we could compute an information transmission figure from the matrix by categorizing the scales. It is not necessary, however, to set up such a categorized matrix, since uncertainties can be estimated from variances (Garner & McGill, 1956) and we can easily determine the appropriate variances with no need to categorize.

Information transmission can be computed from the following equation:

$$U(R{:}S) = U(R) - U_S(R), \quad [2]$$

where $U(R)$ is the response information or uncertainty, and $U_S(R)$ is the response information conditional on the stimulus, or response equivocation, as it is sometimes called.

$$\text{est}U(R{:}S) = \tfrac{1}{2}\log_2 [V(R)] - \tfrac{1}{2}\log_2 [V_S(R)] \quad [3]$$

In this equation $V(R)$ is the variance of the response continuum, and $V_S(R)$ is the average variance of the responses taking one stimulus at a time, i.e., the within column variance. The two terms on the right side of Equation 3 are the estimates of the equivalent uncertainties of Equation 2. (Actually, each of these terms in Equation 3 should have an associated term which reflects the width of the category interval, but since these terms are the same in both cases, they cancel each other—see Garner & McGill, 1956, for further explanation.)

Now we know that the value of $V_S(R)$ is 1.0, since it is so by definition in a unit normalized system; therefore the logarithm of it is zero.

On the other hand,

$$V(R) = V(\bar{R}) + 1 = V(\bar{S}) + 1. \quad [4]$$

$V(\bar{S})$ is the variance of the actual stimulus scale values, but since we are assuming a one-to-one relation between the responses and the stimuli (except for error variation), the mean response for each stimulus is the same as the stimulus value itself. Thus the variance of the *means* of the responses is the same as the variance of the stimulus scale values. But the total variance of the responses, $V(R)$, will be the variance of the means plus the variance within groups (the error variance in analysis of variance terms). Since the error variance is 1.00, the total variance of the responses in this idealized matrix will be $V(\bar{S}) + 1$.

Thus the estimated second term on the right of Equation 2 is zero, and the first term on the right is directly estimated as shown in Equation 1.

The other point of view from which to understand Equation 1 is by means of correlational analysis. Attneave (1959) has pointed out the general relation that

$$\text{est}U(R{:}S) = -\tfrac{1}{2}\log_2(1 - r^2) \quad [5]$$

Now we can define

$$r_s^2 = V(\bar{S})/[V(\bar{S}) + 1] \quad [6]$$

This correlation coefficient is stated in the usual form as the ratio of the predictable variance to the total variance of a set of scores, and as shown above, the total variance of the responses is $V(\bar{S}) + 1$, and the predictable variance of the responses is $V(\bar{S})$. This correlation coefficient is quite similar to one proposed by Gulliksen and Tukey (1958, p. 108), except that their coefficient is a reliability of *mean* scale values, while the coefficient proposed here is essentially a reliability of individual scores, and is a straightforward variance ratio.

From the relation of Equation 6 $(1\text{-}r_s^2)$ is $1/[V(\bar{S}) + 1]$, and from Equation 5, Equation 1 is directly derivable.

Characteristics of the Estimate

This estimating procedure is quite accurate for these data. It should be pointed out, however, that the accuracy of the estimating procedure depends on the normality of the distribution of discriminability scale values, since the estimating equations work accurately, as both Garner and McGill (1956) and Attneave (1959) have pointed out, only when both variances are approximately of the same distribution. This requirement means that the normal distribution is necessary for the mean scale values since the error variance is already forced not only into a unit system but also into a normal system. If the stimulus scale values are off to the extent of providing a rectangular distribution, the error in estimating could be as much as 0.25 bit. The estimating procedure can be used quite generally, however, if a correction factor is estimated for the shape of the distribution.

It is interesting to note that these estimates of information transmission agree very well with the information transmission measures corrected for bias, rather than with the uncorrected measures, even though no unbiasing procedure is inherent in the calculations. The reason for this fact is that the discriminability scales are based on essentially all of the available data, and thus are not themselves subject to the same sorts of bias which operate with the directly computed measures of information transmission. In this respect, the estimating procedure accomplishes much the same effect in minimizing the problem of bias as do the procedures suggested by Attneave

(1955) for pooling data in various ways which are appropriate to specific situations.

THE OPTIMUM NUMBER OF CATEGORIES

The primary purpose of this paper is to demonstrate that information measures and discriminability scales are useful in determining the optimum number of rating categories, and furthermore to show that these two measures are rationally related. It is nevertheless appropriate to discuss the nature of the factors which affect the optimum number of rating categories.

The optimum number of rating categories, or at least the number beyond which there will be no further improvement in discrimination, is clearly a function of the amount of discriminability inherent in the stimuli being rated. Thus there can be no single number of rating categories appropriate to all rating situations. At the limits of discriminability, this fact is obvious. For example, if ratings are made of a set of objects between which there are no perceptible differences, then one rating category will do as well as two or more. If there is no perceptual discrimination, there is no need of a discriminating rating scale.

On the other hand, suppose that we have twenty different objects all of which are clearly discriminable. Then we must have as many rating categories as there are objects in order for raters to demonstrate their ability to discriminate, and ranking will be the most effective procedure. In information terms, we cannot allow the response information to limit the measurable information transmission.

The limiting cases are manifest; the intermediate cases are much more ambiguous. Bendig and Hughes (1953) had their subjects rate 12 countries for familiarity, and when the Miller-Madow correction for bias is applied to their data, there is no improvement in information transmission beyond three rating categories, with which the information transmission obtained is approximately 0.25 bit. When Bendig (1954) had subjects rate preference for 20 foods, increased information transmission (again corrected for bias) was obtained up to at least nine categories, with which 0.32 bit of transmitted information was obtained. In the present experiment, increased discrimination is obtained with up to 20 categories, with which approximately 0.50 bit of transmitted information was measured.

What seems to be clear from these few cases is that the number of useful categories increases with an increase in the amount of information transmission, and that the number of categories needed is far greater than the information transmission figure seems to indicate.

This fact that the number of useful rating categories is far greater than the information transmission seems to indicate is at first glance very surprising, particularly since in the absolute judgment situation, which is so similar, quite another conclusion has been drawn. When information transmission has been measured in the absolute judgment context, the obtained measure has been used to determine the minimum number of categories with which the given information transmission can be obtained, and many experiments have demonstrated the validity of this interpretation of information measures. In other words, it has been demonstrated that there is a channel capacity, such that no further information transmission can be obtained if the number of categories is increased beyond that indicated by the channel capacity.

The absolute judgment paradigm and the rating scale paradigm are in fact, however, quite different, and the essence of the difference is this: In the absolute judgment paradigm, the number of stimuli and the number of responses are varied concomitantly, while by the nature of the problem they cannot be in the usual rating scale application. In the absolute judgment case, if we have an information transmission of 2.0 bits, we can interpret this figure as indicating that we can obtain just as good discrimination if we use four stimuli and four responses as if we use any larger number. Furthermore, we can and should select the particular four stimuli in such a way that maximum discrimination will be realized.

The nature of the rating scale problem, however, is that we have a fixed number of objects or samples to rate, and we are not free to make this number smaller. If 20 different foods, or 50 men in a plant, are to be rated, we are not really interested in seeing how well we can transmit information if fewer foods or fewer men are rated: we want information about each and every stimulus to be rated.

As long as we are not free to select a small representative set of stimuli (i.e., cannot recode into a more efficient set), then we will need to have a large number of rating categories to obtain relatively small amounts of information transmission. If two adjacent stimuli on a discriminability scale are reliably different, then the rating scale must be sufficiently detailed to allow that difference to be shown.

The nature of this problem is easily shown with reference to the hypothetical set of data shown in Table 2. For this table, the contingent uncertainty, or transmitted information, is 1.07 bits. Now we can easily select two stimuli and two responses which will allow us to obtain 1.00 bit of contingent uncertainty—we simply take the first and fourth stimuli and assign one response to each and the pattern of the data makes it clear that we can expect to obtain no confusions between these stimuli. Thus we will obtain 1.00 bit of transmitted information.

TABLE 2

An Illustrative Contingency Table

Responses	Stimuli				
	1	2	3	4	Σ
1	0.21	0.04			0.25
2	0.04	0.17	0.04		0.25
3		0.04	0.17	0.04	0.25
4			0.04	0.21	0.25
Σ	0.25	0.25	0.25	0.25	1.00

Note.—Cell entries are proportions of joint occurrence.

Suppose, however, that we must keep all the stimuli, but want to reduce to two response categories. The best we can do is to group Response 1 with 2 and also 3 with 4. But since the responses to the second and third stimuli will overlap these two new response categories, there must remain some response equivocation, and we cannot obtain the maximum of 1.00 bit of information transmission. In this actual case, if the data from Table 2 are pooled to form the 4 × 2 matrix, the resultant information transmission is 0.68 bit.

Thus the essential difference between the absolute judgment situation and the rating scale situation is not in the character of the stimuli, but rather in the fact that we feel free to select stimuli in one case and do not in the other. If loudness ratings were obtained under equivalent conditions, i.e., if we keep a fixed and large total number of stimuli, but change the number of rating categories, information transmission would vary with the number of rating categories also.

The small amount of available data suggest that the response information (i.e., logarithm to the Base 2 of the number of categories) should be about 5 to 10 times as large as the expected information transmission, at least for small amounts of information transmission. For small numbers of stimuli to be rated, it is almost certainly worth using complete ranking rather than coarser scales. And as a last statement, it is clear that information transmission cannot be lost by increasing the number of rating categories. Therefore, it is better to err on the side of having too many categories than to err by having too few.

SUMMARY

The purpose of this paper is to show how information transmission measures and discriminability scales can be used to determine the optimum number of rating categories. An illustrative set of data were obtained for 20 samples of handwriting with 4, 6, 8, 10, 12, 16, and 20 rating categories. Information transmission measures, after correction for sampling bias, show a small but definite increase in discrimination up to the 20 categories used. Discriminability scale values were also constructed from the rating data, and the standard deviations of the scale values show the same increase in discrimination. The variance of the discriminability scale values can be used to estimate information transmission, and the estimates agree very well with the directly calculated information measures. It is pointed out that the optimum number of rating categories must depend on the discriminability inherent in the particular rated stimuli, and that the optimum number of categories will be large compared to the measured information transmission.

REFERENCES

ATTNEAVE, F. The estimation of transmitted information when conditional probabilities are interdependent. In H. Quastler (Ed.), *Information theory in psychology.* Glencoe, Ill.: Free Press, 1955.

ATTNEAVE, F. *Applications of information theory to psychology.* New York: Holt, 1959.

BENDIG, A. W. Transmitted information and the length of rating scales. *J. exp. Psychol.,* 1954, **47**, 303–308.

BENDIG, A. W., & HUGHES, J. B. Effect of amount of verbal anchoring and number of rating-scale categories upon transmitted information. *J. exp. Psychol.,* 1953, **46**, 87–90.

GARNER, W. R., & HAKE, H. W. The amount of information in absolute judgments. *Psychol. Rev.,* 1951, **58**, 446–459.

GARNER, W. R., & MCGILL, W. J. The relation between information and variance analyses. *Psychometrika,* 1956, **21**, 219–228.

GUILFORD, J. P. The computation of psychological values from judgments in absolute categories. *J. exp. Psychol.,* 1938, **22**, 32–42.

GUILFORD, J. P. *Psychometric methods.* (2nd ed.) New York: McGraw-Hill, 1954.

GULLIKSEN, H., & TUKEY, J. W. Reliability for the law of comparative judgment. *Psychometrika,* 1958, **23**, 95–110.

MILLER, G. A. Note on the bias of information estimates. In H. Quastler (Ed.), *Information theory in psychology.* Glencoe, Ill.: Free Press, 1955.

QUASTLER, H. (Ed.) *Information theory in psychology.* Glencoe, Ill.: Free Press, 1955.

(Received September 17, 1959)

Psychological Review
1960, Vol. 67, No. 6, 353–372

THE MULTIPLE ABSTRACT VARIANCE ANALYSIS EQUATIONS AND SOLUTIONS:

FOR NATURE–NURTURE RESEARCH ON CONTINUOUS VARIABLES

RAYMOND B. CATTELL[1]

University of Illinois

ESSENTIAL NATURE OF THE METHOD

Research in psychological genetics operates either on: (*a*) particulate behavior syndromes, e.g., Huntington's chorea, phenylpyruvic amentia, which are visibly either present or absent, and which can usually be directly handled by Mendelian concepts and analyses applied to familial observations; or, (*b*) apparently continuously distributed behavior traits, the variance on which one can attempt to assign to various environmental and genetic sources. Some inferences in terms of the former can to a slight degree be made from the latter type of biometrical and population genetics observations through the mathematical treatments of Fisher (1930), Haldane (1932), Kempthorne (1957), Mather (1949), Wright (1931), and others. However, psychologists would in general be content, and, indeed, greatly heartened, even to be given some reliable information about the nature–nurture variance ratios for the more important continuous human personality and ability dimensions. For the overwhelming majority of psychological measures are, as far as we can see, continuous, and the instances that can be directly analyzed by particulate Mendelian analysis are negligibly few.

Until a few years ago all published instances of nature–nurture variance analysis ratios for psychological variables were based on studies of identical and fraternal twins (sometimes sibs). At that time the present writer (1953) put forward a design called the Multiple Abstract Variance Analysis method, which seemed capable of answering more questions than the twin method and of avoiding certain artificialities and limitations therein. Although developed independently in the field of psychological measurements, it proved on examination to be much the same in general principle as the estimation of components of variance used in agricultural and biological genetics. However, the MAVA method (as we shall henceforth abbreviate it) contains features unique to the human cultural situation and to psychological measurement. Moreover, it is new to psychologists, being exemplified in only two studies. The first, measuring personality factors by questionnaire, occurred in 1955 (Cattell, Blewett, & Beloff), and the second, using objective, behavioral batteries for the personality factors, and a culture-fair measure of intelligence, was published in 1957 (Cattell, Kristy, & Stice).

---

[1] The writer wishes to express his great indebtedness for discussion and advice to the following (who are nevertheless not to be held responsible for conclusions): geneticists N. E. Morton and W. J. Schull; geneticists and statisticians H. W. Norton and R. W. Touchberry of the University of Illinois; and statisticians D. L. Burkholder, R. A. Fisher, Q. McNemar, and, especially, C. R. Rao of Calcutta. He wishes further to thank his assistants Merlin Foster, John Horn, and Marcia Sedgwick for their persistence and care in the algebraic treatment of the equations.

The results of these first applications, in that they demonstrated diverse and characteristic nature–nurture ratios for well known personality dimensions, agreeing, moreover, with ulterior evidence, were encouraging. It seemed very desirable, therefore, to develop further the potentialities of the method itself.[2] This paper sets out to make a closer examination of the method, its assumptions, its calculations, and its needs in terms of more refined development.

The title of the method derives from the fact that *conceptual* variances are introduced, i.e., variances which can only be inferred or "abstracted" from actual, observed variances. The label designates this abstractness in order to avoid confusion with the usual use of the analysis of variance design in which questions are asked purely in regard to concrete variances. It is "multiple" because it views each concrete observed variance as the result of several, abstract, inferred variances. The MAVA method in fact postulates that the observed variance on a given trait, e.g., a personality factor score, measurable on a continuum, can be resolved into four contributory variances as follows:

Variance through causes resident in the differences of environment *between* families. Henceforth written σ^2_{be}.

Variance through causes resident in the differences of heredity *between* families. Henceforth written σ^2_{bh}.

Variance through causes resident in the differences of environment for individuals *within* the family. Henceforth, σ^2_{we}.

Variance through causes resident in the differences of heredity between siblings *within* the same family (gene segregation). Henceforth, σ^2_{wh}.

Furthermore, it does not assume that these sources are uncorrelated but attempts simultaneously to determine what correlations exist between these hereditary and environmental variances, both within families (henceforth, $r_{we.wh}$), between families (henceforth, $r_{be.bh}$), and, occasionally, through the rare correlation of a within and between variance. Obviously, the equations below imply that heredity and environment act mutually in a simple additive fashion, *without* any "catch all" category to accommodate to all kinds of interaction effects, as usually conceived in analysis of variance. At the present stage of the model, one could either assume heredity and environment to be uncorrelated and then permit interaction, as in analysis of variance, or sacrifice interaction. The assumption of uncorrelated heredity and environment seemed to us, in the psychological field, less tenable.[3]

A more ideal model would be one which both allows covariance, as here, and also interaction, but, while such a model is being worked out, the present model can be expected to give the usual approximate fit of a simpler model to a, possibly existent, more complex natural pattern. Additionally, the assumption will be recognized that the values obtained for the variances and correlations are specifically those existing in *the given cultural and racial sample* (referable, if desired, to parameters for a statistical population of that type), and that generalization beyond this, e.g., to modern, primitive, etc. cultures, is possible only in the light of additional, strategic, experiments. Since a complete statement of assumptions in a model can be endless, the reader should accept those obviously implied, while others will be mentioned later as they become relevant.

---

[2] The writer has to apologize for five years of delay in presenting this fuller statement, occasioned by nothing more than writing difficulties in presenting a simultaneously adequate explanation of genetic concepts, for psychological readers, and of psychological measurement assumptions, for geneticists, in anything less than book length. The delay is regrettable because psychological genetic researches have been started in the interim on what is essentially an obsolete twin design, or researchers have been compelled laboriously to derive independently from the initial statement (1953) the formula now given in Tables 1 and 2. This situation may perhaps justify the present necessarily uncomfortably condensed presentation.

[3] Were it not that one sometimes hears the remarkable statement that "psychological traits cannot be wholly resolved into genetic and environmental causes" it might be unnecessary to add here that we assume that the variances and covariances are *exhaustive* of the measured trait variances.

The general manner of building equations relating observed to abstract variances will be evident from inspecting, say, Equation 4 in Table 1 below. Here it is stated that the observed variance within a series of pairs of siblings reared together, σ^2_{ST}, is the sum of within family environmental variance, σ^2_{we}, plus within family heredity variance (due to different segregations of parental genes), σ^2_{wh}, plus the covariance (if these influences are correlated), $2r\sigma_{we}\sigma_{wh}$. The calculation of the *observed* variances, as such, will be discussed later. Originally only six such equations were written (Cattell, 1953), because it was presumed that there would only be six unknowns—four variances and two correlation coefficients. Thus by the solution of six simultaneous equations, using the six experimentally determinable concrete variances—σ^2_{ST}, σ^2_{SA}, etc.—it was hoped to solve for the six abstract, conceptual values.

However, six equations and six observed variances are not enough, because: (*a*) The last two or three of the six equations import other unknowns. For example, one of our necessary equations was for twins reared together and, unless one makes the psychologically unsatisfying assumption that the typical environmental difference in the lives of twins is the same as that between sibs, a new term, for a special kind of intersib environmental difference, $\sigma^2_{we''}$, must be introduced. (*b*) Similarly it turns out that with additional experimental groups more than the above *two* heredity–environment covariances are involved. (*c*) Linear dependencies systematically appear between the equations for "within," "between," and "total" variances in each of the family types one uses, thus reducing the number of effective equations.

The additional equations that may be used, and the alternative assumptions possible in using various combinations will be discussed *after* the following section, clarifying the calculation of concrete variances.

CALCULATIONS AND ASSUMPTIONS CONCERNING CONCRETE, EXPERIMENTALLY MEASURED VARIANCES

Conceptually, the within family variance, on some such trait as intelligence, surgency, etc., is the variance (*a*) among the children, i.e., not including the parents; (*b*) of the typical American family (or whatever cultural or ethnic group one studies); (*c*) affected by size of family only statistically, i.e., to the extent that the variance estimate uses different degrees of freedom with different family sizes (actually we took mostly *pairs* but occasionally trios, from which to estimate this variance); (*d*) corrected for age differences, i.e., it is the variance that would exist, on traits susceptible to age change, if all sibs were brought to the same age; (*e*) corrected for experimental error of measurement, but *not* for function fluctuation.

The justification for using the concept thus is partly experimental practicability and partly the achievement of maximum negotiability in further theoretical or calculation use of the obtained variances. Objection can be raised to (*c*) that large families might have, for example, smaller variance, on certain traits, due to more powerful group "coercion to the bio-social mean" (Cattell, 1957). This can be investigated by our design, through an experiment with a sample of small size families contrasted with one of large size. Meanwhile the adopted assumption is simpler. As to (*d*), it is true that by allowing an age correction we are already here using an "abstracted" rather than a literal, concrete variance; but there is no point in confounding psychological variance with age variance. If one knows the age variance of families in a given culture, the concrete variances can always be readily "restored" from the obtained variances through this more analytical approach and, meanwhile, the nature–nurture ratio is more pure, comparable, and comprehensible for an age-constant group.

The issue in (*e*) is best seen by setting out the equation:

$$\underset{\substack{\text{True}\\\text{(Individual Difference)}\\\text{Variance}}}{\sigma^2_t} = \underset{\substack{\text{Observed}\\\text{Variance}}}{\sigma^2_o} - \underset{\substack{\text{Error of}\\\text{Measurement}}}{\sigma^2_e} - \underset{\substack{\text{Function}\\\text{Fluctuation}}}{\sigma^2_f} \qquad [1]$$

Although the allowance for error of measurement, derived from the *dependability* coefficient of the test (Cattell, 1957) by $\sigma^2_e = 1 - r_{rd}$, is always only an estimate, it is better than not allowing for it at all. The question whether function fluctuation should be taken out depends on one's conception of environment. Periodic or aperiodic, short-interval, function fluctuation occurs in most personality and ability measures (Cattell, 1957; Fiske & Rice, 1955; Thouless, 1936) from internal or external causes. Should one say that the ergs of hunger or sleep or the anxiety reaction are not innate because they fluctuate hourly? The writer is inclined to the view that over a sufficient series of standard situations a reasonably stable strength level could be assigned to an individual on such fluctuant traits and that it is *this* level which we speak about when we talk of individual differences, e.g., in anxiety. For most purposes it is the nature–nurture contribution to this relatively stable "over a couple of months" value, at various age levels, that we are interested in. Function fluctuation variance is calculable from the difference of the dependability and stability coefficients (Cattell, 1957), but, for lack of the latter it was not actually taken out of the variance in our experiments (Cattell, 1957; Cattell, Blewett, & Beloff, 1955). When not taken out it theoretically swells the environmental variance, and it is significant that the distribution of ratios for two dozen personality dimensions in our studies was decidedly skewed to the environmental predominance end. If experimental error is not taken out, it also predominantly swells the variance ascribed to environment. Ideally, therefore, both should be taken out before calculating the nature–nurture ratios.

The assumptions and problems in the estimation of the observed between-family variance are essentially of the same nature. Again, one is estimating a value from the sample for a statistical population assumed to be homogeneous with the sample, and, in this case, the correctness of (*b*) above—that the statistical population is the typical American population—is harder to justify in the circumstances commonly prevailing in nature–nurture studies. For, without quite extraordinary exertions by the experimenters, a disproportionate number of cases will come from social welfare agencies. Obviously, the ideal requirement is a *k*-way classification of families by religion, occupation, age of parents, economic, social, education, etc., background of parents, etc., corresponding to a stratified population sample (or, when investigating the effect of particular conditions, e.g., religion, on the ratios, a deliberately selected sample).

Another specification which must be watched in defining and using the various variance ratios from the MAVA is whether the differences considered are between boy, girl,[4] or boy-and-girl families. Even in the third case the equation for identical twins must, naturally, remain that for same sex pairs; but, provided $\sigma^2_{we''}$ has been used for the special between-twin environmental variance these same sex pairs can be used in the total solution.

As to the degrees of freedom, the *within-family* values, when pairs of children are taken (deviation of each from their common mean, squared and added over k families), will be k. When n children instead of two are taken from each family it will be $k(n - 1)$. For *between-family* variance, obtained by subtracting the grand mean of all (sib) families from the mean of each family, squaring and adding, it will be $k - 1$.

CONCEPTS AND ASSUMPTIONS IN THE EQUATIONS

It will be evident from the initial illustration by the case of within-family environmental variance above that the experimenter in this area is placed in a theoretical dilemma peculiarly embarrassing to a perfectionist. As he adds more equations in the hope of more completely determining unknowns, he adds new unknowns. In the writer's opinion one can achieve a complete solution in the sense that the number of experimental situations capable of yielding new equations eventually overhauls the number of unknowns. But since it is not beyond the wit of man to invent even finer conceptual subdivisions of family types and situations to which he would like a separate answer, let us admit that the process is asymptotic. Consequently, it is up to the experimenter to draw a line where he will in data gathering, in the light of present estimations of magnitudes of various influences and the steps dictated by research strategy.

Accordingly, the presentation is cut to two designs, one, called the Limited Resources Design, which makes grosser assumptions in order to accommodate to the resources of nine out of ten investigators, and the other, called the Complete Design, which aims at what, by ordinary standards, can be called a complete solution, regardless of difficulty or expense. The assumptions in these two designs can best be discussed when we have considered the symbols used, as follows:

Abstract Variances

1. σ^2_{we} = trait variance (between sibs) due to within-family environmental (*we*) variation. Incidentally, we assume that this variability of environment among children raised in the same family is essentially the same for all kinds of families, including normal and adoptive families. On the other hand in the special case of fraternal twins, who are of identical age, and identical twins, who, according to clinicians, have an urge to be more

---

[4] Not only environmental, but also hereditary within family variances could be different for girl-girl pairs, due to the extra chromosome.

alike or (in other opinions) more individualistic, we enter with special terms capable of discovering such differences, as follows:

2. $\sigma^2_{we'}$ = trait variance within family associated with type of environmental difference peculiar to fraternal twins.
3. $\sigma^2_{we''}$ = trait variance within family associated with type of environmental difference peculiar to identical twins.
4. σ^2_{wh} = trait variance due to within-family heredity (*wh*) variation.
5. σ^2_{be} = trait variance due to between-family environment (*be*) variation.
6. σ^2_{bh} = trait variance due to between-family hereditary (*bh*) variation.

Another possibility—that the within-family environmental variance among sibs might be different in natural and adoptive families—has not been given separate representation at this point.

Correlations

7. $r_{wh \cdot we}$ = correlation between effects of within family heredity and within family environment (siblings).

A systematically different value for this correlation might arise in special circumstances, notably (*a*) when the two sibs are of the same age, i.e., fraternal twins. For example, a trait which might be fostered in a younger by an older brother might be inhibited in the interests of competition when they are the same age; (*b*) when it is an adoptive family, i.e., not true sibs and raised by foster parents. For, whether we believe with G. B. Shaw that parents are allergic to their own hereditary defects appearing in their children, or take a more kindly view that they will be more understanding, the $r_{we.wh}$ will at least be different. Moreover, the hereditary variance in adoptive (e.g., foster) families is typically greater than for ordinary families, being a sum (or other combination) of within and between hereditary variances. If there are reasons for r settling down to a certain value with a certain variance, it is unlikely to remain at the same value when the range of one variable is altered. Actually we have assumed here that increased variance of heredity (Equation 6, Table 1) and of environment (Equation 5, Table 1) affect the $r_{wh.we}$ term in the same essential fashion and we have represented both by a single new r as follows:

8. $r'_{wh \cdot we}$ = correlation of heredity and environment in adoptive families.

In such families also, it will be noted, we have the peculiar condition of a possible correlation between a between-group and a within-group variance, in that the foster parents may, for example, treat the children differently according to the hereditary variations appearing from the families from which they derive. Hence:

9. $r'_{we \cdot bh}$ = correlation of within-environment and between-heredity peculiar to adoptive families. A prime is placed on this r, though there is no other $r_{we \cdot bh}$ from which to distinguish it, because it helps classification to have all r's peculiar to adoptive families marked with a prime.
10. $r_{bh \cdot be}$ = correlation between effects of between-family heredity and between-family environment. The frequently discussed correlation of "good" heredity with "good" environment among families is a possible instance.

There follow two possible correlations, probably nonzero only because of what may be called "placement agency" effects. In our society this arises partly from the deliberate policy of welfare agencies of placing more able children in families likely to give them a better educational environment, but in most societies it might arise from a tendency to adopt orphaned *relatives*. These placement agency r's are marked by a double prime, as follows:

11. $r''_{wh \cdot be}$ = placement agency correlation peculiar to sibs raised in different families, where within-family hereditary differences may become correlated with between-family environment by deliberate intention of social agencies in placing, e.g., brighter sib in more educated family.
12. $r''_{bh \cdot be}$ = placement agency correlation, through any systematic tendency to match biological family characters of adopted children to environmental characters of foster parent families.

Between four main variables—*we*, *wh*, *be*, *bh*—there are six possible correlations (more, with special variables above) of which four would normally be dropped on the assumption that no correlation commonly exists of within and between variances. However, having found two exceptions we should look for more. The possibility that $r_{wh.bh}$ and $r_{we.be}$ could be nonzero is however, hard to conceive. For example, it is difficult to see any way in which an individual's deviation in heredity from his family mean could be connected with the fact that his family differs in mean from the community hereditary mean. On the other hand, if, say, specific "roles" are more emphasized in highly than poorly educated families, the possibility exists that an individual's deviation environmentally from his family mean will be greater when his family is environmentally most positively deviant from the community mean. This would give a significant curvilinear correlation but no significant linear correlation. This will be recognized as another restriction on our model, that it does not respond to curvilinear correlation among variances. But it does mean that no especial reason exists for complicating our equations with an $r_{be.we}$ term. We are left, therefore, with two "main" (r) correlations, two "component" (r') correlations, and two "placement" (r'') correlations requiring terms for solutions.

If space permitted, a number of other interesting theoretical possibilities could be discussed here, though most would be rejected as highly improbable. What may seem an undue refinement of conception, but is nevertheless vital in later interpretations and applications, is implicit in the above labeling of within and between-family hereditary and environmental variances as "effects" not "influences." It would not be correct, for example, rigidly to reason that all environmental variance within families comes from influences within the family or all environmental variance between families from influences within the community operating differently on these families. Similarly the correlation of a within-family hereditary deviation with a within-family environmental effect does not imply that the correlation is brought about entirely by influences within the family. Tracing influences is a second step after locating effects.

To determine the 12 unknowns above (6 variances, 6 r's), which we have found it necessary to introduce on theoretical grounds in the Complete Design, one must employ, for reasons stated, more than 12 equations. The following 16 experimentally determinable variances, each for a particular kind of group, are required to represent the left side of each of the equations needed. (The equations themselves are set out in Tables 1 and 2).

1. σ^2_{ITT} = variance in the trait as measured among identical twins reared together by their parents.
2. σ^2_{ITA} = variance in the trait as measured among identical twins reared apart.
3. σ^2_{FTT} = variance in the trait as measured among fraternal twins

raised together by their true parents.

4. σ^2_{ST} = variance in the trait as measured among ordinary siblings reared together by their parents.
5. σ^2_{SA} = variance in the trait as measured among ordinary siblings reared apart.
6. σ^2_{UT} = variance in the trait as measured among unrelated children reared together in one family with foster parents who are biological parents of none.
7. σ^2_{UA} = variance in the trait as measured among unrelated reared apart (general population).
8. σ^2_{BITTF} = variance in the trait as measured between the means of identical twin pairs in their true families.
9. σ^2_{BITAF} = variance in the trait as measured between the means of identical twin pairs raised apart.
10. σ^2_{BNF} = variance in the trait as measured between the means of natural families, i.e., means of sib pairs (or trios, etc.) reared together.
11. σ^2_{BBF} = variance in the trait as measured between the means of biological families whose siblings are reared in different foster homes.
12. σ^2_{BSF} = variance in the trait as measured between the means of social families, i.e., adoptive families of biologically unrelated children.
13. σ^2_{HST} = variance in the trait as measured among half-siblings raised together.
14. σ^2_{HSA} = variance of half-siblings raised apart.
15. σ^2_{BHSTF}[5] = variance in the trait as measured among means between half-siblings raised together by one true parent.
16. σ^2_{BHSAF} = variance of means between half-siblings raised apart.

Other familial groups yielding experimentally determinable variances are theoretically conceivable, but they demand too much tact on the part of the experimenter in locating them, or too much moral aberration on the part of the community, to be practicable!

EQUATIONS FOR THE COMPLETE SOLUTION AND THE LIMITED RESOURCES DESIGNS

The 16 equations which link the above experimental, concrete variances with the abstract variances and correlations seemed to produce them are shown in Table 1 below. We assume they will require no discussion, beyond the illustration already given of a typical equation (Equation 4) for most readers. Perhaps a few words on a more esoteric example—Equation 13, half-sibs reared together—will complete the explanation.

Here we might have two children of one mother by two fathers. Incidentally, introduction of a different term for children of two mothers by one father is not considered necessary at this stage, nor is the variance considered modified by the different heads of family in which the two half-sibs might be brought up together. They will differ by (*a*) within-family heredity variance, σ^2_{wh}. For each is only one "choice" among the possible parental gene segregations; (*b*) also, in this case, by half the interfamily hereditary variance, $\frac{1}{2}\sigma^2_{bh}$, for one parent is the same and one different; and (*c*) by the usual within-family environmental differences, σ^2_{we}. In terms of covariances there will be (*d*) a "component variance" correlation of the between-family hereditary variance, (which is now part of the total within-family environmental variance) with environmental variance in the family. This produces the covariance term, $\sqrt{2}r'_{we.bh}\sigma_{we}\sigma_{bh}$. (The coefficient $\sqrt{2}$ before this is derived from $2/\sqrt{2}$, the usual coefficient

---

[5] A further specialization can be made here according to whether child is raised with mother or father.

TABLE 1

THE SEVENTEEN GENERALLY PRACTICABLE EQUATIONS FOR THE COMPLETE MULTIPLE VARIANCE DESIGN

Note.—The arrangement of equations in Table 1 can, for purposes of simplification be regarded as creating two blocks: 1. 1 through 12, which are constellations of ordinary sib and twin families, and which divide into *within* variances (1 through 6) and *between* variances (8 through 12), and, 2. constellations in half-sib families, again the first half being *within* and the second *between* variances. The matching of each *between* with its corresponding *within* is indicated by (Suppl. "X")—Supplementary to the earlier numbered equation—written beside the matched equation.

$$\sigma^2_{ITT} = \sigma^2_{we''} \tag*{[1]}$$

$$\sigma^2_{ITA} = \sigma^2_{we} + \sigma^2_{be} \tag*{[2]}$$

$$\sigma^2_{FTT} = \sigma^2_{wh} + \sigma^2_{we'} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we'} \qquad \text{(or replace last term by } 2r_{wh \cdot we'}\sigma_{wh}\sigma_{we'}\text{)} \tag*{[3]}$$

$$\sigma^2_{ST} = \sigma^2_{wh} + \sigma^2_{we} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} \tag*{[4]}$$

$$\sigma^2_{SA} = \sigma^2_{wh} + \sigma^2_{we} + \sigma^2_{be} + 2r'_{wh \cdot we}\sigma_{wh}\sigma_{we}[+ 2r''_{wh \cdot be}\sigma_{wh}\sigma_{be}] \tag*{[5]}$$

$$\sigma^2_{UT} = \sigma^2_{wh} + \sigma^2_{we} + \sigma^2_{bh} + 2r'_{wh \cdot we}\sigma_{wh}\sigma_{we} + 2r'_{we \cdot bh}\sigma_{we}\sigma_{bh} \tag*{[6]}$$

$$\sigma^2_{UA} = \sigma^2_{wh} + \sigma^2_{we} + \sigma^2_{bh} + \sigma^2_{be} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} + 2r_{bh \cdot be}\sigma_{bh}\sigma_{be} \tag*{[7]}$$

(Suppl. 1)

$$\sigma^2_{BITTF} = 2\sigma^2_{wh} + \sigma^2_{we} + 2\sigma^2_{bh} + 2\sigma^2_{be} + 2\sqrt{2}r_{wh \cdot we}\sigma_{wh}\sigma_{we} + 4r_{bh \cdot be}\sigma_{bh}\sigma_{be} \tag*{[8]}$$

(Suppl. 2)

$$\begin{aligned}\sigma^2_{BITAF} = {} & 2\sigma^2_{wh} + \sigma^2_{we} + 2\sigma^2_{bh} + \sigma^2_{be} + 2\sqrt{2}r'_{wh \cdot we}\sigma_{wh}\sigma_{we} + 2\sqrt{2}r''_{bh \cdot be}\sigma_{bh}\sigma_{be} \\ & + 2\sqrt{2}r'_{we \cdot bh}\sigma_{we}\sigma_{bh}[+ 2\sqrt{2}r''_{wh \cdot be}\sigma_{wh}\sigma_{be}]\end{aligned} \tag*{[9]}$$

$$[\sigma^2_{BFTT} = \sigma^2_{wh} + \sigma^2_{we'''} + 2\sigma^2_{bh} + 2\sigma^2_{be} + 2r_{wh \cdot we'''}\sigma_{wh}\sigma_{we'''} + 4r_{bh \cdot be}\sigma_{bh}\sigma_{be}] \tag*{[9a]}$$

(Suppl. 4 or 3 and 4)

$$\sigma^2_{BNF} = \sigma^2_{wh} + \sigma^2_{we} + 2\sigma^2_{bh} + 2\sigma^2_{be} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} + 4r_{bh \cdot be}\sigma_{bh}\sigma_{be} \tag*{[10]}$$

(Suppl. 5)

$$\begin{aligned}\sigma^2_{BBF} = {} & \sigma^2_{wh} + \sigma^2_{we} + 2\sigma^2_{bh} + \sigma^2_{be} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} + 2\sqrt{2}r'_{we \cdot bh}\sigma_{we}\sigma_{bh} \\ & + 2\sqrt{2}r''_{bh \cdot be}\sigma_{bh}\sigma_{be}[+ 2r''_{wh \cdot be}\sigma_{wh}\sigma_{be}]\end{aligned} \tag*{[11]}$$

(Suppl. 6)

$$\begin{aligned}\sigma^2_{BSF} = {} & \sigma^2_{wh} + \sigma^2_{we} + \sigma^2_{bh} + 2\sigma^2_{be} + 2r'_{wh \cdot we}\sigma_{wh}\sigma_{we} + 2r'_{we \cdot bh}\sigma_{we}\sigma_{bh} \\ & + 2\sqrt{2}r''_{bh \cdot be}\sigma_{bh}\sigma_{be}[+ 2\sqrt{2}r''_{wh \cdot be}\sigma_{wh}\sigma_{be}]\end{aligned} \tag*{[12]}$$

$$\sigma^2_{HST} = \sigma^2_{wh} + \sigma^2_{we} + \tfrac{1}{2}\sigma^2_{bh} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} + \sqrt{2}r'_{we \cdot bh}\sigma_{we}\sigma_{bh} \tag*{[13]}$$

$$\sigma^2_{HSA} = \sigma^2_{wh} + \sigma^2_{we} + \tfrac{1}{2}\sigma^2_{bh} + \sigma^2_{be} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} + \sqrt{2}r_{bh \cdot be}\sigma_{bh}\sigma_{be} \tag*{[14]}$$

(Suppl. 13)

$$\sigma^2_{BHSTF} = \sigma^2_{wh} + \sigma^2_{we} + \tfrac{3}{2}\sigma^2_{bh} + 2\sigma^2_{be} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} + 2\sqrt{3}r_{bh \cdot be}\sigma_{bh}\sigma_{be} \tag*{[15]}$$

(Suppl. 14)

$$\sigma^2_{BHSAF} = \sigma^2_{wh} + \sigma^2_{we} + \tfrac{3}{2}\sigma^2_{bh} + \sigma^2_{be} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} + \sqrt{6}r_{bh \cdot be}\sigma_{bh}\sigma_{be} \tag*{[16]}$$

2 being multiplied by the root of a half because σ^2_{bh} is present only as a *half* variance.) (*e*) The usual covariance of heredity and environment, $2r_{we.wh}\sigma_{we}\sigma_{wh}$. Other equations are similarly derived and from consideration of each it will be evident that a good deal of careful reasoning from first principles, and discussion among experts, has gone into them.

TABLE 2

THE TEN EQUATIONS USED IN THE LIMITED RESOURCES DESIGN

$$\sigma^2_{ITT} = \sigma^2_{we'} \qquad [1]$$

$$\sigma^2_{FTT} = \sigma^2_{wh} + \sigma^2_{we'} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we'} \qquad [2]$$

$$\sigma^2_{ST} = \sigma^2_{wh} + \sigma^2_{we} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} \qquad [3]$$

$$\sigma^2_{SA} = \sigma^2_{wh} + \sigma^2_{we} + \sigma^2_{be} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} \qquad [4]$$

$$\sigma^2_{UT} = \sigma^2_{wh} + \sigma^2_{we} + \sigma^2_{bh} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} + 2r'_{we \cdot bh}\sigma_{we}\sigma_{bh} \qquad [5]$$

$$\sigma^2_{UA} = \sigma^2_{wh} + \sigma^2_{we} + \sigma^2_{bh} + \sigma^2_{be} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} + 2r_{bh \cdot be}\sigma_{bh}\sigma_{be} \qquad [6]$$

$$\sigma^2_{BITTF} = 2\sigma^2_{wh} + \sigma^2_{we} + 2\sigma^2_{bh} + 2\sigma^2_{be} + 2\sqrt{2}r_{wh \cdot we}\sigma_{wh}\sigma_{we} + 4r_{bh \cdot be}\sigma_{bh}\sigma_{be} \qquad [7]$$

$$\sigma^2_{BNF} = \sigma^2_{wh} + \sigma^2_{we} + 2\sigma^2_{bh} + 2\sigma^2_{be} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} + 4r_{bh \cdot be}\sigma_{bh}\sigma_{be} \qquad [8]$$

$$\sigma^2_{BBF} = \sigma^2_{wh} + \sigma^2_{we} + 2\sigma^2_{bh} + \sigma^2_{be} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} + 2\sqrt{2}r'_{we \cdot bh}\sigma_{we}\sigma_{bh} + 2\sqrt{2}r''_{bh \cdot be}\sigma_{bh}\sigma_{be} \qquad [9]$$

$$\sigma^2_{BSF} = \sigma^2_{wh} + \sigma^2_{we} + \sigma^2_{bh} + 2\sigma^2_{be} + 2r_{wh \cdot we}\sigma_{wh}\sigma_{we} + 2r'_{we \cdot bh}\sigma_{we}\sigma_{bh} + 2\sqrt{2}r''_{bh \cdot be}\sigma_{bh}\sigma_{be} \qquad [10]$$

To get the systematic organization of terms needed in such a presentation we have followed the rule of putting heredity terms before environmental ones, variance before covariance, within-family before between-family. The primes and double primes on r's are explained above. Occasionally, where a mathematically *possible* combinatory term might also seem *needed* by the reader, but has been dropped by us on fuller consideration, it has been placed in brackets, i.e., strictly we regard the term as unnecessary. We have deliberately refrained from working out as decimals the various numerical coefficients, since their representation as radicals enables the reader more easily to perceive their derivation.

The equations of Table 1 are presented as a totality of experimentally available equations, from which an experimenter may draw according to the research resources available to him and the degree of simplification of assumptions which he cares to make. Even the Complete Design cannot simultaneously use all of them, because of the linear dependencies noted above, though this does not avoid collecting the data that would be required to use virtually all of them.[6] The 11 equations which we advocate as most effective for the final determinations of all 11 unknowns in 16 are: 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, and 14. The only regrettable waste in this choice is that it does not use sibs raised together, Equation 4, an easily obtainable group, but the alternatives are worse. For example, half-sibs raised together are needed in most others, and these are socially difficult to identify, with scientific reliability.

The costly features of the Complete Design are identical twins reared apart and half-sibs reared apart, both of which require much effort to locate. Accordingly we have suggested for most researchers the Limited Resources Design in Table 2, which uses only 10 equations and requires only sibs, and, in twins, those reared together. A sufficient sample of these should be locatable by any investigator near a large city. It is not possible to create this design simply by lifting 10 equations from Table 1; for the assumptions that are made in each have to be changed in order to be able to reduce the total number of equations. The simpler assumptions, (*a*) consider the environmental variance for identical and fraternal twins raised together to be identical, as $\sigma^2_{we'}$; (*b*) omit differentiation of $r'_{wh.we}$ and $r_{wh.we}$; i.e., the modification of the environment–heredity correlation when the hereditary and the environmental variance are increased in families, respectively, of unrelated children raised together and sibs raised apart. Thus it solves for five variances and three correlations, instead of for six variances and five correlations as in the Complete Design.

---

[6] Commonly, one has the choice of using either the within or the between variance from data gathered on a particular type of familial group.

THE SOLUTIONS AND THEIR CONFIDENCE LIMITS

Since the solutions of the above equations are a considerable undertaking—mainly because of the need to choose the best sequences in combining them—it would be helpful to have tables set out for the general solutions. The results of some three or four months of algebraic work and checking are set out in Tables 3 and 4, in which the researcher can simply insert his experimentally-obtained, concrete variances to obtain the abstract variances. In order not to interrupt the main presentation these have been placed at the end of the article. Quite apart from the substantial algebraic task there are certain theoretical problems in solving the equations of Table 1, which have been discussed in another context by Nelder (1953). Some of these are approached below.

Additionally to the solutions in Tables 3 and 4, which are only as reliable as human pursuit of long manipulations of simultaneous equations can be, and should accordingly be checked before use, solutions have been worked out along different lines on the electronic computer. A matrix, A, consisting of the coefficients for variances and covariances arranged with rows corresponding to the 11 equations from the Complete Design, is post-multiplied by a Vector X of abstract variances and covariances to yield the vector of concrete variances, B, thus

$$AX = B$$

Solving for X would give the abstract variances, thus,

$$X = A^{-1}B$$

But, since we wish to compare the coefficients for the concrete variances in these solutions with those in Tables 3 and 4, which have to be utilized in the following discussion on standard errors and confidence limits, we stopped the solution short at obtaining A^{-1}. This was done by the following transformation because the Limited Resources Design, unlike the Complete Design, does not present a square matrix of coefficients

$$A^{-1} = (A'A)^{-1}A'$$

These matrices of coefficients may be obtained by application to the author.

The principal incompleteness of the MAVA method as first applied (Cattell, Blewett, & Beloff, 1955) to psychological data was the lack of standard errors and confidence limits for the values obtained. As Tables 3 and 4 remind one, each of the abstract variances finally obtained is a function of quite a number of experimentally-determined variances, while the correlation coefficients have at least an equally complex derivation.

At the present juncture in research it becomes important to attempt to solve the complex problem here presented. An additional complexity

TABLE 3

SOLUTIONS FOR SIX ABSTRACT VARIANCES AND FIVE HEREDITY-ENVIRONMENT CORRELATIONS, IN THE COMPLETE DESIGN

Note.—For convenience, certain equations below, namely, those having a quadratic solution, are set out in triplicate. The first statement shows the roots which give possible positive and negative entries, while the following "a" and "b" equations give the actual resulting alternative answers—and in decimals since there is no longer need to reveal the mode of derivation. In these cases both solutions are retained until further considerations cause one to be ruled out.

$$\sigma^2_{we} = 2\sigma^2_{UA} - .5\sqrt{2}\sigma^2_{BITTF} - \sigma^2_{SA} + .5\sqrt{2}\sigma^2_{BITAF} - .5\sqrt{2}\sigma^2_{BBF} + \sigma^2_{ITA} + (.5\sqrt{2}-1)\sigma^2_{BNF} \qquad [1]$$

$$\sigma^2_{we} = 2\sigma^2_{UA} - .7070\sigma^2_{BITTF} - \sigma^2_{SA} + .7070\sigma^2_{BITAF} - .7070\sigma^2_{BBF} + \sigma^2_{ITA} - .2930\sigma^2_{BNF} \qquad [1a]$$

$$\sigma^2_{we} = 2\sigma^2_{UA} + .7070\sigma^2_{BITTF} - \sigma^2_{SA} - .7070\sigma^2_{BITAF} + .7070\sigma^2_{BBF} + \sigma^2_{ITA} - 1.7070\sigma^2_{BNF} \qquad [1b]$$

$$\sigma^2_{we'} = 2r^2_{wh\cdot we}\sigma^2_{wh} - (\sigma^2_{wh} - \sigma^2_{FTT}) - r_{wh\cdot we}\sqrt{4r^2_{wh\cdot we}\sigma^4_{wh} - 4\sigma^2_{wh} + 4\sigma^2_{wh}\sigma^2_{FTT}} \qquad [2]^a$$

$$\sigma^2_{we''} = \sigma^2_{ITT} \qquad [3]$$

$$\sigma^2_{wh} = (.5+.5\sqrt{2})\sigma^2_{BITTF} - .5\sqrt{2}\sigma^2_{SA} + .5\sigma^2_{BITAF} - .5\sigma^2_{BBF} - (.5+.5\sqrt{2})\sigma^2_{BNF} + .5\sqrt{2}\sigma^2_{ITA} \qquad [4]$$

$$\sigma^2_{wh} = 1.2070\sigma^2_{BITTF} - .7070\sigma^2_{SA} + .5\sigma^2_{BITAF} - .5\sigma^2_{BBF} - 1.2070\sigma^2_{BNF} + .7070\sigma^2_{ITA} \qquad [4a]$$

$$\sigma^2_{wh} = -.2070\sigma^2_{BITTF} + .7070\sigma^2_{SA} + .5\sigma^2_{BITAF} - .5\sigma^2_{BBF} + .2070\sigma^2_{BNF} - .7070\sigma^2_{ITA} \qquad [4b]$$

$$\sigma^2_{be} = .5\sqrt{2}\sigma^2_{BITTF} + \sigma^2_{SA} - 2\sigma^2_{UA} - .5\sqrt{2}\sigma^2_{BITAF} + .5\sqrt{2}\sigma^2_{BBF} + (1-.5\sqrt{2})\sigma^2_{BNF} \qquad [5]$$

$$\sigma^2_{be} = .7070\sigma^2_{BITTF} + \sigma^2_{SA} - 2\sigma^2_{UA} - .7070\sigma^2_{BITAF} + .7070\sigma^2_{BBF} + .2930\sigma^2_{BNF} \qquad [5a]$$

$$\sigma^2_{be} = -.7070\sigma^2_{BITTF} + \sigma^2_{SA} - 2\sigma^2_{UA} + .7070\sigma^2_{BITAF} - .7070\sigma^2_{BBF} + 1.7070\sigma^2_{BNF} \qquad [5b]$$

$$\sigma^2_{bh} = \sigma^2_{BITTF} + \sqrt{2}\sigma^2_{SA} + (2+\sqrt{2})\sigma^2_{UA} - \sigma^2_{BITAF} + \sigma^2_{BBF} - \sigma^2_{BNF} - (2\sqrt{2}+2)\sigma^2_{HSA} \qquad [6]$$

$$\sigma^2_{bh} = \sigma^2_{BITTF} + 1.4141\sigma^2_{SA} + 3.4141\sigma^2_{UA} - \sigma^2_{BITAF} + \sigma^2_{BBF} - \sigma^2_{BNF} - 4.8282\sigma^2_{HSA} \qquad [6a]$$

$$\sigma^2_{bh} = \sigma^2_{BITTF} - 1.4141\sigma^2_{SA} + .5859\sigma^2_{UA} - \sigma^2_{BITAF} + \sigma^2_{BBF} - \sigma^2_{BNF} + .8282\sigma^2_{HSA} \qquad [6b]$$

$$r_{wh\cdot we} = \frac{\begin{gathered}.25\sigma^2_{BNF} - (.25\sqrt{2}+.5)\sigma^2_{ITA} - .25\sigma^2_{BITTF} + (.25\sqrt{2}+.5)\sigma^2_{SA}\\ - (.25\sqrt{2}+.25)\sigma^2_{BITAF} + (.25\sqrt{2}+.25)\sigma^2_{BBF}\end{gathered}}{\begin{gathered}\sqrt{(.5+.5\sqrt{2})\sigma^2_{BITTF} - .5\sqrt{2}\sigma^2_{SA} + .5\sigma^2_{BITAF} - .5\sigma^2_{BBF} - (.5+.5\sqrt{2})\sigma^2_{BNF} + .5\sqrt{2}\sigma^2_{ITA}}\\ \times\sqrt{2\sigma^2_{UA} - .5\sqrt{2}\sigma^2_{BITTF} - \sigma^2_{SA} + .5\sqrt{2}\sigma^2_{BITAF} - .5\sqrt{2}\sigma^2_{BBF} + \sigma^2_{ITA} + (.5\sqrt{2}-1)\sigma^2_{BNF}}\end{gathered}} \qquad [7]$$

$$r'_{wh\cdot we} = \frac{\begin{gathered}(.25\sqrt{2}+.5)\sigma^2_{SA} - (.25\sqrt{2}+.5)\sigma^2_{BSF} + (.25\sqrt{2}+.25)\sigma^2_{BNF} - (.25\sqrt{2}+.25)\sigma^2_{BITTF}\\ - .25\sigma^2_{BITAF} + .25\sigma^2_{BBF}\end{gathered}}{\begin{gathered}\sqrt{(.5+.5\sqrt{2})\sigma^2_{BITTF} - .5\sqrt{2}\sigma^2_{SA} + .5\sigma^2_{BITAF} - .5\sigma^2_{BBF} - (.5+.5\sqrt{2})\sigma^2_{BNF} + .5\sqrt{2}\sigma^2_{ITA}}\\ \times\sqrt{2\sigma^2_{UA} - .5\sqrt{2}\sigma^2_{BITTF} - \sigma^2_{SA} + .5\sqrt{2}\sigma^2_{BITAF} - .5\sqrt{2}\sigma^2_{BBF} + \sigma^2_{ITA} + (.5\sqrt{2}-1)\sigma^2_{BNF}}\end{gathered}} \qquad [8]$$

$$r'_{we\cdot bh} = \frac{\begin{gathered}(\sqrt{2}+1)\sigma^2_{HSA} + (1-.25\sqrt{2})\sigma^2_{BNF} - .5\sqrt{2}\sigma^2_{SA} - (.2+.5\sqrt{2})\sigma^2_{UA}\\ + (.5-.25\sqrt{2})\sigma^2_{BITAF} + (.25\sqrt{2}-.5)\sigma^2_{BITTF} + (.25\sqrt{2}-.5)\sigma^2_{BBF} + .5\sigma^2_{UT}\end{gathered}}{\begin{gathered}\sqrt{2\sigma^2_{UA} - .5\sqrt{2}\sigma^2_{BITTF} - \sigma^2_{SA} + .5\sqrt{2}\sigma^2_{BITAF} - .5\sqrt{2}\sigma^2_{BBF} + \sigma^2_{ITA} + (.5\sqrt{2}-1)\sigma^2_{BNF}}\\ \times\sqrt{\sigma^2_{BITTF} + 2\sigma^2_{SA} + (2+\sqrt{2})\sigma^2_{UA} - \sigma^2_{BITAF} + \sigma^2_{BBF} - \sigma^2_{BNF} - (2\sqrt{2}+2)\sigma^2_{HSA}}\end{gathered}} \qquad [9]$$

---

[a] For the sake of simplification it is best to determine values for $r_{wh.we}$ and σ^2_{wh} and substitute these in this equation to determine $\sigma^2_{we'}$.

[b] Solutions 12 through 16 are not directly for any required end value—there are only 11 real unknowns. They are introduced as conveniences in the process of solution. They consist of phrases containing radicals, and were treated as single unknowns in the solution of the original equations. The correlations $r_{wh.we}$ etc., can be directly obtained for these solutions by dividing the solution obtained by the product of the corresponding squares of the variances indicated on the left-hand side of the equation. Or, if it is desired the four solutions (7–11) may be used and by direct substitution, the values for $r_{wh.we}$, etc., can be obtained. In Equations 7 through 11, only one solution is given to save time and space. (If the plus value of 2 is used in obtaining the variances (Equations 1–6) consistency must be followed by using plus values throughout.

TABLE 3—*Continued*

$$\begin{aligned} r_{bh\cdot be}=&\frac{(.5+.25\sqrt{2})\sigma^2_{BITAF}-(.5+.25\sqrt{2})\sigma^2_{BITTF}-(.5+.5\sqrt{2})\sigma^2_{UA}-(.5+.25\sqrt{2})\sigma^2_{BBF}}{\sqrt{.5\sqrt{2}\sigma^2_{BITTF}+\sigma^2_{SA}-2\sigma^2_{UA}-.5\sqrt{2}\sigma^2_{BITAF}+.5\sqrt{2}\sigma^2_{BBF}+(1-.5\sqrt{2})\sigma^2_{BNF}}}\\ &\frac{-(.5+.5\sqrt{2})\sigma^2_{SA}+(\sqrt{2}+1)\sigma^2_{HSA}+(.5+.25\sqrt{2})\sigma^2_{BNF}}{\times\sqrt{\sigma^2_{BITTF}+\sqrt{2}\sigma^2_{SA}+(2+\sqrt{2})\sigma^2_{UA}-\sigma^2_{BITAF}+\sigma^2_{BBF}-\sigma^2_{BNF}-(2\sqrt{2}+2)\sigma^2_{HSA}}}\end{aligned}\qquad [10]$$

$$\begin{aligned} r''_{bh\cdot be}=&\frac{\sigma^2_{HSA}+(.75\sqrt{2}-.75)\sigma^2_{BNF}+(.25\sqrt{2}-1)\sigma^2_{SA}+(1-.5\sqrt{2})\sigma^2_{UA}+(.75\sqrt{2}-.25)\sigma^2_{BITAF}}{\sqrt{.5\sqrt{2}\sigma^2_{BITTF}+\sigma^2_{SA}-2\sigma^2_{UA}-.5\sqrt{2}\sigma^2_{BITAF}+.5\sqrt{2}\sigma^2_{BBF}+(1-.5\sqrt{2})\sigma^2_{BNF}}}\\ &\frac{+(.25-.75\sqrt{2})\sigma^2_{BITTF}+(.25-.5\sqrt{2})\sigma^2_{BBF}-.5\sigma^2_{UT}}{\times\sqrt{\sigma^2_{BITTF}+\sqrt{2}\sigma^2_{SA}+(2+\sqrt{2})\sigma^2_{UA}-\sigma^2_{BITAF}+\sigma^2_{BBF}-\sigma^2_{BNF}-(2\sqrt{2}+2)\sigma^2_{HSA}}}\end{aligned}\qquad [11]$$

$$\begin{aligned} r_{wh\cdot we}\sigma_{wh}\sigma_{we}=&.25\sigma^2_{BNF}-(.25\sqrt{2}+.5)\sigma^2_{ITA}-.25\sigma^2_{BITTF}+(.25\sqrt{2}+.5)\sigma^2_{SA}\\ &-(.25\sqrt{2}+.25)\sigma^2_{BITAF}+(.25\sqrt{2}+.25)\sigma^2_{BBF}\end{aligned}\qquad [12]^{b}$$

$$r_{wh\cdot we}\sigma_{wh}\sigma_{we}=.25\sigma^2_{BNF}-.8535\sigma^2_{ITA}-.25\sigma^2_{BITTF}+.8535\sigma^2_{SA}-.8535\sigma^2_{BITAF}+.8535\sigma^2_{BBF}\qquad [12a]$$

$$r_{wh\cdot we}\sigma_{wh}\sigma_{we}=.25\sigma^2_{BNF}-.1465\sigma^2_{ITA}-.25\sigma^2_{BITTF}+.1465\sigma^2_{SA}-.1465\sigma^2_{BITAF}+.1465\sigma^2_{BBF}\qquad [12b]$$

$$\begin{aligned} r'_{wh\cdot we}\sigma_{wh}\sigma_{we}=&(.25\sqrt{2}+.5)\sigma^2_{SA}-(.25\sqrt{2}+.5)\sigma^2_{BSF}+(.25\sqrt{2}+.25)\sigma^2_{BNF}\\ &-(.25\sqrt{2}+.25)\sigma^2_{BITTF}-.25\sigma^2_{BITAF}+.25\sigma^2_{BBF}\end{aligned}\qquad [13]$$

$$r'_{wh\cdot we}\sigma_{wh}\sigma_{we}=.8535\sigma^2_{SA}-.8535\sigma^2_{BSF}+.6035\sigma^2_{BNF}-.6035\sigma^2_{BITTF}-.25\sigma^2_{BITAF}+.25\sigma^2_{BBF}\qquad [13a]$$

$$r'_{wh\cdot we}\sigma_{wh}\sigma_{we}=.1465\sigma^2_{SA}-.1465\sigma^2_{BSF}-.1035\sigma^2_{BNF}+.1035\sigma^2_{BITTF}-.25\sigma^2_{BITAF}+.25\sigma^2_{BBF}\qquad [13b]$$

$$\begin{aligned} r'_{we\cdot bh}\sigma_{we}\sigma_{bh}=&(\sqrt{2}+1)\sigma^2_{HSA}+(1-.25\sqrt{2})\sigma^2_{BNF}-.5\sqrt{2}\sigma^2_{SA}-(2+.5\sqrt{2})\sigma^2_{UA}\\ &+(.5-.25\sqrt{2})\sigma^2_{BITAF}+(.25\sqrt{2}-.5)\sigma^2_{BITTF}+(.25\sqrt{2}-.5)\sigma^2_{BBF}+.5\sigma^2_{UT}\end{aligned}\qquad [14]$$

$$\begin{aligned} r'_{we\cdot bh}\sigma_{we}\sigma_{bh}=&2.4141\sigma^2_{HSA}+.6565\sigma^2_{BNF}-.7070\sigma^2_{SA}-2.7070\sigma^2_{UA}+.1465\sigma^2_{BITAF}\\ &-.1465\sigma^2_{BITTF}-.1465\sigma^2_{BBF}+.5\sigma^2_{UT}\end{aligned}\qquad [14a]$$

$$\begin{aligned} r'_{we\cdot bh}\sigma_{we}\sigma_{bh}=&-.4141\sigma^2_{HSA}+1.3535\sigma^2_{BNF}+.7070\sigma^2_{SA}-1.2930\sigma^2_{UA}+.8535\sigma^2_{BITAF}\\ &-.8535\sigma^2_{BITTF}-.8535\sigma^2_{BBF}+.5\sigma^2_{UT}\end{aligned}\qquad [14b]$$

$$\begin{aligned} r_{bh\cdot be}\sigma_{bh}\sigma_{be}=&(.5+.25\sqrt{2})\sigma^2_{BITAF}-(.5+.25\sqrt{2})\sigma^2_{BITTF}-(.5+.5\sqrt{2})\sigma^2_{UA}\\ &-(.5+.25\sqrt{2})\sigma^2_{BBF}-(.5+.5\sqrt{2})\sigma^2_{SA}+(\sqrt{2}+1)\sigma^2_{HSA}+(.5+.25\sqrt{2})\sigma^2_{BNF}\end{aligned}\qquad [15]$$

$$\begin{aligned} r_{bh\cdot be}\sigma_{bh}\sigma_{be}=&.8535\sigma^2_{BITAF}-.8535\sigma^2_{BITTF}-1.2070\sigma^2_{UA}-.8535\sigma^2_{BBF}-1.2070\sigma^2_{SA}\\ &+2.4141\sigma^2_{HSA}+.8535\sigma^2_{BNF}\end{aligned}\qquad [15a]$$

$$\begin{aligned} r_{bh\cdot be}\sigma_{bh}\sigma_{be}=&+.1465\sigma^2_{BITAF}-.1465\sigma^2_{BITTF}+.2070\sigma^2_{UA}-.1465\sigma^2_{BBF}+.2070\sigma^2_{SA}\\ &-.4141\sigma^2_{HSA}+.4650\sigma^2_{BNF}\end{aligned}\qquad [15b]$$

$$\begin{aligned} r''_{bh\cdot be}\sigma_{bh}\sigma_{be}=&\sigma^2_{HSA}+(.75\sqrt{2}-.75)\sigma^2_{BNF}+(.25\sqrt{2}-1)\sigma^2_{SA}+(1-.5\sqrt{2})\sigma^2_{UA}\\ &+(.75\sqrt{2}-.25)\sigma^2_{BITAF}+(.25-.75\sqrt{2})\sigma^2_{BITTF}+(.25-.5\sqrt{2})\sigma^2_{BBF}-.5\sigma^2_{UT}\end{aligned}\qquad [16]$$

$$\begin{aligned} r''_{bh\cdot be}\sigma_{bh}\sigma_{be}=&\sigma^2_{HSA}+.3106\sigma^2_{BNF}-.6465\sigma^2_{SA}+.2930\sigma^2_{UA}+.8106\sigma^2_{BITAF}\\ &-.8106\sigma^2_{BITTF}-.4570\sigma^2_{BBF}-.5\sigma^2_{UT}\end{aligned}\qquad [16a]$$

$$\begin{aligned} r''_{bh\cdot be}\sigma_{bh}\sigma_{be}=&\sigma^2_{HSA}-1.8106\sigma^2_{BNF}-1.3535\sigma^2_{SA}+1.7070\sigma^2_{UA}-1.3106\sigma^2_{BITAF}\\ &+1.3106\sigma^2_{BITTF}+.9570\sigma^2_{BBF}-.5\sigma^2_{UT}\end{aligned}\qquad [16b]$$

lies in the fact that most investigators (including the present writer), have usually not been content with the values for the four main variances but have proceeded to calculate various ratios between these, or between certain sums of these, e.g., the ratio of *total* environmental to *total* hereditary variance. For these final *ratios* are more likely to express the ultimate

TABLE 4

SOLUTIONS FOR THE FIVE VARIANCES AND THREE CORRELATIONS IN THE LIMITED RESOURCES DESIGN

$$\sigma^2_{we}=(2-\sqrt{2})\sigma^2_{ST}+\sigma^2_{SA}-\sigma^2_{UT}+(2\sqrt{2}-2)\sigma^2_{UA}-\sigma^2_{BITTF}+(2-\sqrt{2})\sigma^2_{BNF}+\sigma^2_{BBF}-\sigma^2_{BSF} \tag*{[1]^a}$$

$$\sigma^2_{we}=.5859\sigma^2_{ST}+\sigma^2_{SA}-\sigma^2_{UT}+.8282\sigma^2_{UA}-\sigma^2_{BITTF}+.5859\sigma^2_{BNF}+\sigma^2_{BBF}-\sigma^2_{BSF} \tag*{[1a]}$$

$$\sigma^2_{we}=3.4141\sigma^2_{ST}+\sigma^2_{SA}-\sigma^2_{UT}-4.8181\sigma^2_{UA}-\sigma^2_{BITTF}+3.4141\sigma^2_{BNF}+\sigma^2_{BBF}-\sigma^2_{BSF} \tag*{[1b]}$$

$$\sigma^2_{we'}=\sigma^2_{ITT} \tag*{[2]}$$

$$\begin{aligned}\sigma^2_{wh}=&\,\sigma^2_{ST}(\sqrt{2}-1)+.5\sqrt{2}\sigma^2_{SA}-.5\sqrt{2}\sigma^2_{UT}+(2-2\sqrt{2})\sigma^2_{UA}+\sigma^2_{BITTF}\\&+(\sqrt{2}-2)\sigma^2_{BNF}+.5\sqrt{2}\sigma^2_{BBF}-.5\sqrt{2}\sigma^2_{BSF}\end{aligned} \tag*{[3]}$$

$$\begin{aligned}\sigma^2_{wh}=&\,.4141\sigma^2_{ST}+.7070\sigma^2_{SA}-.7070\sigma^2_{UT}-.8282\sigma^2_{UA}+\sigma^2_{BITTF}-.5859\sigma^2_{BNF}\\&+.7070\sigma^2_{BBF}-.7070\sigma^2_{BSF}\end{aligned} \tag*{[3a]}$$

$$\begin{aligned}\sigma^2_{wh}=&-2.4141\sigma^2_{ST}-.7070\sigma^2_{SA}+.7070\sigma^2_{UT}+4.8282\sigma^2_{UA}+\sigma^2_{BITTF}\\&-3.4141\sigma^2_{BNF}-.7070\sigma^2_{BBF}+.7070\sigma^2_{BSF}\end{aligned} \tag*{[3b]}$$

$$\sigma^2_{be}=\sigma^2_{SA}-2\sigma^2_{UA}+\sigma^2_{BNF} \tag*{[4]}$$

$$\sigma^2_{bh}=(1+.5\sqrt{2})\sigma^2_{SA}+(1+.5\sqrt{2})\sigma^2_{BBF}-2\sigma^2_{UA}+\sigma^2_{BNF}-(1+.5\sqrt{2})\sigma^2_{BSF}-.5\sqrt{2}\sigma^2_{UT} \tag*{[5]}$$

$$\sigma^2_{bh}=1.7070\sigma^2_{SA}+1.7070\sigma^2_{BBF}-2\sigma^2_{UA}+\sigma^2_{BNF}-1.7070\sigma^2_{BSF}-.7070\sigma^2_{UT} \tag*{[5a]}$$

$$\sigma^2_{bh}=.2930^2_{SA}+.2930\sigma^2_{BBF}-2\sigma^2_{UA}+\sigma^2_{BNF}-.2930\sigma^2_{BSF}+.7070\sigma^2_{UT} \tag*{[5b]}$$

$$r_{wh\cdot we}=\frac{.5\sigma^2_{FTT}-.5\sigma^2_{ITT}-.5\sigma^2_{BITTF}-.25\sqrt{2}\sigma^2_{SA}-.25\sqrt{2}\sigma^2_{BBF}}{\sqrt{\sigma^2_{SA}(\sqrt{2}-1)+.5\sqrt{2}\sigma^2_{SA}-.5\sqrt{2}\sigma^2_{UT}+(2-2\sqrt{2})\sigma^2_{UA}+\sigma^2_{BITTF}}}$$

(Square root sign continued)

$$\frac{+.25\sqrt{2}\sigma^2_{BSF}+(1-.5\sqrt{2})\sigma^2_{BNF}+(\sqrt{2}-1)\sigma^2_{UA}}{+(\sqrt{2}-2)\sigma^2_{BNF}+.5\sqrt{2}\sigma^2_{BBF}-.5\sqrt{2}\sigma^2_{BSF}}$$

$$\frac{+(.5-.5\sqrt{2})\sigma^2_{ST}+(.25\sqrt{2})\sigma^2_{UT}}{\times\sqrt{(2-\sqrt{2})\sigma^2_{ST}+\sigma^2_{SA}-\sigma^2_{UT}+(2\sqrt{2}-2)\sigma^2_{UA}-\sigma^2_{BITTF}+(2-\sqrt{2})\sigma^2_{BNF}+\sigma^2_{BBF}-\sigma^2_{BSF}}} \tag*{[6]}$$

$$r'_{we\cdot bh}=\frac{.25\sigma^2_{UT}-.25\sigma^2_{SA}}{\sqrt{(2-\sqrt{2})\sigma^2_{ST}+\sigma^2_{SA}-\sigma^2_{UT}+(2\sqrt{2}-2)\sigma^2_{UA}-\sigma^2_{BITTF}+(2-\sqrt{2})\sigma^2_{BNF}+\sigma^2_{BBF}-\sigma^2_{BSF}}}$$

$$\frac{-.25\sigma^2_{BBF}+.25\sigma^2_{BSF}}{\times\sqrt{(1+.5\sqrt{2})\sigma^2_{SA}+(1+.5\sqrt{2})\sigma^2_{BBF}-2\sigma^2_{UA}+\sigma^2_{BNF}-(1+.5\sqrt{2})\sigma^2_{BSF}-.5\sqrt{2}\sigma^2_{UT}}} \tag*{[7]}$$

$$r_{bh\cdot be}=\frac{.25\sqrt{2}\sigma^2_{UT}-(1+.25\sqrt{2})\sigma^2_{SA}+1.5\sigma^2_{UA}-.5\sigma^2_{BNF}}{\sqrt{\sigma^2_{SA}-2\sigma^2_{UA}+\sigma^2_{BNF}}}$$

$$\frac{-(.25\sqrt{2}+.5)\sigma^2_{BBF}+(.25\sqrt{2}+.5)\sigma^2_{BSF}}{\times\sqrt{(1+.5\sqrt{2})\sigma^2_{SA}+(1+.5\sqrt{2})\sigma^2_{BBF}-2\sigma^2_{UA}+\sigma^2_{BNF}-(1+.5\sqrt{2})\sigma^2_{BSF}-.5\sqrt{2}\sigma^2_{UT}}} \tag*{[8]}$$

$$r''_{bh\cdot be}=\frac{\sqrt{2}\sigma^2_{UA}-.5\sqrt{2}\sigma^2_{SA}-.5\sqrt{2}\sigma^2_{BNF}}{\sqrt{\sigma^2_{SA}-2\sigma^2_{UA}+\sigma^2_{BNF}}}$$

---

[a] These variances (Subequations a and b) are given solutions wherein $\sqrt{2}$ assumes a value of 1.4141 in 1a throughout, and wherein $\sqrt{2}$ assumes a value of -1.4141 in 1b.

[b] Solutions 10 through 14 are for phrases containing radicals. These phrases were treated, as in Table 1, as single unknowns in the solution of the original equations. The correlations $r_{wh.we}$ etc., can be directly obtained from these solutions by dividing the solution obtained by the product of the corresponding squares of the variances indicated on the left-hand side of the equation. Or, if it is desired the four solutions (6–9) may be used and by direct substitution, the values for $r_{wh.we}$, etc., can be obtained. In Equations 6 through 9, only one solution is given to save time and space. (If the plus values of 2 are used in obtaining the variances (Equations 1–5) consistency must be followed by using plus values throughout.)

TABLE 4—*Continued*

$$\frac{+.25\sqrt{2}\sigma^2_{BSF}-.25\sqrt{2}\sigma^2_{UT}}{\times\sqrt{(1+.5\sqrt{2})\sigma^2_{SA}+(1+.5\sqrt{2})\sigma^2_{BBF}-2\sigma^2_{UA}+\sigma^2_{BNF}-(1+.5\sqrt{2})\sigma^2_{BSF}-.5\sqrt{2}\sigma^2_{UT}}} \qquad [9]$$

$$\begin{aligned} r_{wh\cdot we}\sigma_{wh}\sigma_{we'} = {} & .5\sigma^2_{FTT}-.5\sigma^2_{ITT}-.5\sigma^2_{BITTF}-.25\sqrt{2}\sigma^2_{SA}-.25\sqrt{2}\sigma^2_{BBF}+.25\sqrt{2}\sigma^2_{BSF} \\ & +(1-.5\sqrt{2})\sigma^2_{BNF}+(\sqrt{2}-1)\sigma^2_{UA}+(.5-.5\sqrt{2})\sigma^2_{ST}+.25\sqrt{2}\sigma^2_{UT} \qquad [10]^{b} \end{aligned}$$

$$\begin{aligned} r_{wh\cdot we}\sigma_{wh}\sigma_{we'} = {} & .5\sigma^2_{FTT}-.5\sigma^2_{ITT}-.5\sigma^2_{BITTF}-.3535\sigma^2_{SA}-.3535\sigma^2_{BBF}+.3535\sigma^2_{BSF} \\ & +.2930\sigma^2_{BNF}+.4141\sigma^2_{UA}-.2070\sigma^2_{ST}+.3535\sigma^2_{UT} \qquad [10a] \end{aligned}$$

$$\begin{aligned} r_{wh\cdot we}\sigma_{wh}\sigma_{we'} = {} & .5\sigma^2_{FTT}-.5\sigma^2_{ITT}-.5\sigma^2_{BITTF}+.3535\sigma^2_{SA}+.3535\sigma^2_{BBF}-.3535\sigma^2_{BSF} \\ & +1.7070\sigma^2_{BNF}-2.4141\sigma^2_{UA}+1.2070\sigma^2_{ST}-.3535\sigma^2_{UT} \qquad [10b] \end{aligned}$$

$$\begin{aligned} r_{wh\cdot we}\sigma_{wh}\sigma_{we} = {} & -.5\sigma^2_{ST}+(.5+.25\sqrt{2})\sigma^2_{UT}-(.5+.25\sqrt{2})\sigma^2_{SA}-(.5+.25\sqrt{2})\sigma^2_{BBF} \\ & +(.5+.25\sqrt{2})\sigma^2_{BSF}-.5\sigma^2_{BNF}+\sigma^2_{UA} \qquad [11] \end{aligned}$$

$$r_{wh\cdot we}\sigma_{wh}\sigma_{we} = -.5\sigma^2_{ST}+.8535\sigma^2_{UT}-.8535\sigma^2_{SA}-.8535\sigma^2_{BBF}+.8535\sigma^2_{BSF}-.5\sigma^2_{BNF}+\sigma^2_{UA} \qquad [11a]$$

$$r_{wh\cdot we}\sigma_{wh}\sigma_{we} = -.5\sigma^2_{ST}+.1465\sigma^2_{UT}-.1465\sigma^2_{SA}-.1465\sigma^2_{BBF}+.1465\sigma^2_{BSF}-.5\sigma^2_{BNF}+\sigma^2_{UA} \qquad [11b]$$

$$r'_{we\cdot bh}\sigma_{we}\sigma_{bh} = .25\sigma^2_{UT} - .25\sigma^2_{SA}-.25\sigma^2_{BBF}+.25\sigma^2_{BSF} \qquad [12]$$

$$\begin{aligned} r_{bh\cdot be}\sigma_{bh}\sigma_{be} = {} & .25\sqrt{2}\sigma^2_{UT}-(1+.25\sqrt{2})\sigma^2_{SA}+1.5\sigma^2_{UA}-.5\sigma^2_{BNF} \\ & -(.25\sqrt{2}+.5)\sigma^2_{BBF}+(.25\sqrt{2}+.5)\sigma^2_{BSF} \qquad [13] \end{aligned}$$

$$r_{bh\cdot be}\sigma_{bh}\sigma_{be} = .3535\sigma^2_{UT}-1.3535\sigma^2_{SA}+1.5\sigma^2_{UA}-.5\sigma^2_{BNF}-.8535\sigma^2_{BBF}+.8535\sigma^2_{BSF} \qquad [13a]$$

$$r_{bh\cdot be}\sigma_{bh}\sigma_{be} = -.3535\sigma^2_{UT}-.6465\sigma^2_{SA}+1.5\sigma^2_{UA}-.5\sigma^2_{BNF}-.1465\sigma^2_{BBF}+.1465\sigma^2_{BSF} \qquad [13b]$$

$$r''_{bh\cdot be}\sigma_{bh}\sigma_{be} = \sqrt{2}\sigma^2_{UA}-.5\sqrt{2}\sigma^2_{SA}-.5\sqrt{2}\sigma^2_{BNF}+.25\sqrt{2}\sigma^2_{BSF}-.25\sqrt{2}\sigma^2_{UT} \qquad [14]$$

$$r''_{bh\cdot be}\sigma_{bh}\sigma_{be} = 1.4141\sigma^2_{UA}-.7070\sigma^2_{SA}-.7070\sigma^2_{BNF}+.3535\sigma^2_{BSF}-.3535\sigma^2_{UT} \qquad [14a]$$

$$r''_{bh\cdot be}\sigma_{bh}\sigma_{be} = -1.4141\sigma^2_{UA}+.7070\sigma^2_{SA}+.7070\sigma^2_{BNF}-.3535\sigma^2_{BSF}+.3535\sigma^2_{UT} \qquad [14b]$$

inferences in which one is interested, as a starting point for further theoretical psychological developments and experiments. Yet determining the distribution and the confidence limits of ratios presents an additional challenge to statistical inference.

The typical form of any solution for an abstract variance, as in Tables 3 and 4 above, may be written:

$$\hat{\sigma}^2_a = k_1\sigma^2_1 + k_2\sigma^2_2 \cdots + k_n\sigma^2_n \qquad [2]$$

where $\hat{\sigma}^2_a$ is the estimated abstract variance, e.g., within family environment; k_1, k_2, etc., are numerical, constants; and σ^2_1, etc., are the estimated concrete variances for particular family relation population species. Our need is to determine the variance of the variance on the left, and to do so it is necessary to decide whether the variances on the right are correlated. Usually within and between mean squares are obviously uncorrelated for any one particular analysis. In this case, since they are quite different groups of subjects (*S*s), it might be argued that the chances of systematic correlation are still less. However, the question is actually a very complex one, to be answered only by study of the structure of the data in each case. For example, in some instances of sibling pairs, they are *not* random subgroups from the larger group. The concensus of expert statistical opinion consulted nevertheless is that, in terms of possible correlation over a series of samples, we are justified in treating these as uncorrelated variances. Accordingly the variance of the abstract variance estimate is a simple sum of the independent functions of the experimentally obtained variance estimates on the right of our equations, i.e.,

$$\hat{\sigma}^2_{\sigma_a^2} = \frac{2\Sigma k_i^2\sigma_i^4}{n_i - 1} \qquad [3]$$

where i is any subgroup.

As stated above, it is likely in most actual experiments, unless one is prepared to throw away hard won data, that the numbers in the various experimentally obtained family constellation subgroups (sibs, twins, etc.) will differ. If the successive groups in Equation 1 have n_1, n_2, etc. cases the above estimate of the variance of the variance is stated more completely as:

$$\hat{\sigma}^2_{\sigma_a^2} = \frac{2k_1^2\sigma_1^4}{n_1 - 1} + \frac{2k_2^2\sigma_2^4}{n_2 - 1} \cdots + \frac{2k_n^2\sigma_n^4}{n_n - 1} \qquad [4]$$

However, if n is small—less than 30 cases (unlikely in these experiments)—the degrees of freedom which give an unbiased estimate become $n_i + 1$, etc., according to Daniels (1939) and Welch (1956).

Thus the confidence limits for a given abstract variance can be designated as:

$$Pr\left[\sigma_a^2 - Z_j\frac{\sqrt{2\Sigma k_i\sigma_i^2}}{n_i - 1} < \hat{\sigma}_a^2 < \sigma_a^2 + Z_j\frac{\sqrt{2\Sigma k_i\sigma_i^2}}{n_i - 1}\right] = P_j \qquad [5]$$

where Z_j is the normal standard deviate for a given probability, P_j; σ_a^2 the "true" value for the population variance, and we assume that, through large n or the use of an unbiased estimate, the distribution of the variance is normal.

Finally, to determine the confidence limits of heredity-environment variance ratio's, e.g., $\frac{\sigma_{bh}^2}{\sigma_{be}^2}$, or, in general $\frac{\sigma_a^2}{\sigma_b^2}$, we simplify the probability statements for numerator and denominator as follows:

$$Pr[A < \hat{\sigma}_a^2 < B] = P_1 \qquad [6a]$$

$$Pr[C < \hat{\sigma}_b^2 < D] = P_2 \qquad [6b]$$

The combined probability for the ratio is not strictly the product P_1P_2 because P_1 and P_2 are not entirely independent; but when using the extreme confidence interval ranges (.95 to .99) in which we are alone interested, the error from assumption of independence is very small[7] and we may use:

$$Pr\left[\frac{A}{B} < \frac{\hat{\sigma}_a^2}{\hat{\sigma}_b^2} < \frac{B}{C}\right] = P_1P_2 \qquad [7]$$

As with any confidence limits test, it is not possible to decide, without actual empirical variance data carrying experimental error, what the desirable number of cases should be in a well planned experiment. However, with such a major experimental undertaking as is involved in a really well designed MAVA experiment, to determine nature–nurture ratios for some particular primary personality source trait, obtaining the earliest possible indication from experience as to what is adequate is vital. Fortunately, a lead can now be obtained by applying the above formulae to the sample sizes achieved in the first investigations of inheritance of primary personality factors by Cattell, Blewett, and Beloff (1955) and Cattell, Kristy, and Stice (1957). These studies measured a total of 1,024 eleven- and twelve-year old children on all personality factors established up to that time, using, first, questionnaire measures (the High School Personality Questionnaire) and, secondly, objective tests (the Children's Objective-Analytic Battery). Taking for illustration first a factor from the questionaire, and one with the highest of hereditary determinations (next to intelligence), we choose H factor, parmia vs. threctia, a temperament factor believed associated with parasympathetic system predominance. The value $\frac{\sigma_{we}^2}{\sigma_{wh}^2}$ came out at $\frac{.73}{1.39} = .52$, making heredity about twice as important, within the family as environment. According to our present formulae above the $P = .05$ limits are approximately .18 and 1.39.

---

[7] For example with both P_1 and P_2 equal to .90 and assuming independence we get $P_1P_2 = .81$, whereas assuming lack of independence this becomes .80.

Taking as different a case as possible for our second illustration we turn to a predominantly environmentally determined factor and one measured in objective tests. Let us examine the between family variance. For this we chose UI 23 (Cattell, 1957) which Eysenck has called the general neuroticism factor and which we call ergic regression. The value of $\frac{\sigma_{be}^2}{\sigma_{bh}^2}$ came out at 6.7, and the $P = .05$ limits by the above formulae stand at 1.9 and 32.0. All values make environment more important than heredity, but the range in this case is very large.

It is well known that as one extends the number of linear combinations from which estimations are made the magnitude of the standard error increases in the disheartening fashion seen in the above results. However, the conclusion to be drawn, since knowledge of the relative environmental and hereditary contributions on continuous personality and ability dimensions is of great theoretical and practical importance (and since the twin method has similarly large sampling errors and is additionally, fundamentally amiss in its assumptions!) is surely that research must foot the bill. To say that the method is self-defeating because it requires 10,000 rather than 1,000 cases is to use an argument which, applied to the analogous wearisome organization required in separating radium, would have brought the work of the Curies to a standstill, or, in our own generation, would have blocked the way to atomic energy. Those who believe that advance in the human sciences is as important as in the physical sciences will not blink at the expenditure which, though large compared with a rat experiment, is still a hundredth of that in several current physical science problems. At present (unless one pursues the mirage of analogy from animal experimentation) the MAVA method, despite its experimental-organizational arduousness, is the only way either to check basic theories or, what is more important, generate more exact and potent hypotheses, as discussed below. Nevertheless, it is presented here as unfinished in development, and requiring improvements, notably in refinement of the estimation of standard errors, in the use of the attenuation correction, and in regard to the assumption of noncorrelation in Equation 7, etc. These issues would require separate papers in themselves.

While no space can be given here to heading off beforehand the numerous misunderstandings possible on so complex a topic, we would point out that perusal of the discussion of assumptions in their psychological context in the earlier articles (Cattell, 1953; Cattell, Blewett, & Beloff, 1955; Cattell, Kristy, & Stice, 1957) will make the present statements of formulae, etc., more meaningful to the psychologist.

PSYCHOLOGICAL THEORY IMPLICIT IN THE MAVA MODEL

What has been propounded above is not a theory, in the sense of most discussed in this *Review*, but a model Burt and Howard (1957), and a few others on intelligence, and of Cattell, Beloff, and Blewett (1955), and Cattell, Kristy, and Stice (1957) on personality, nothing existed to restrict and a method, by which a considerable array of theories can be expressed and tested. In this area the vital need of psychology has been the initial appearance of some dependable relationships out of which a theory worthy of that name, rather than mere speculation, can be born. Apart from the work of Burks (1928), Burt and Howard (1957), and a few others on intelligence, and of Cattell, Beloff, and Blewett (1955), and Cattell, Kristy, and Stice (1957) on personality, nothing existed to restrict speculators from ascribing anything from 0.1% to 99.9% of the variance in any human trait to whichever influence their prejudices favored.

The restrictions which the present model puts upon theories (additive hereditary and environmental effects, interaction of heredity and environment by correlation and covariance but not statistical interaction) are slight compared to the aid it gives in checking many theories. Among the theories deserving experimental investigation are:

1. The position adopted in our own investigations that the variance ratio for any factorially composite piece of behavior will be middling but that more clear-cut predominance of heredity or environment will be found for factor-pure source traits (Cattell, 1946). Although it is theoretically possible, as mooted by Thomson and the present writer (1946) that a genetic factor (gene) may appear as a statistical factor it is almost certain that constitutional factors among the psychologists factors will be polygenic. Nevertheless there is a tendency for factors to be predominantly either environmental or hereditary (Cattell, Kristy, & Stice, 1957).

2. The theory that the occurrence of environmental influences at particular (early) ages, as in Freudian theory or in imprinting (Hess, 1959), is more effective than at others. In animals this theory can be tested by manipulative experiment, but in humans only by some form of the MAVA method applied to groups with different histories. In this connection an editorial critic has asked, "How can one possibly separate maturation from learning, age decline, etc., when these operate differently at different ages?" The answer is that it is precisely through, and only through, first obtaining these ratios for strategically chosen age levels, cultures, etc., that further inferences can be made regarding the periods of life, or the subcultural influences, determining the relative action of maturation and learning. For example, it was only through the discovery of the predominantly negative value of the heredity–environment *correlations*, from the above experiments, that the "law of coercion to the bio-social mean" (see 6 below) became justified as a worthwhile proposition regarding heredity–environment *interaction*.

3. The theory that, contrary to popular thinking, the ratio of hereditary influence will be higher with older *S*s. This assumes that younger children will have been exposed to a smaller sample of the total environment and will experience in this exposure the greater variance of means characteristic of smaller samples. The

hereditary variance remaining unchanged, the nature–nurture ratio should therefore be greater when older *S*s are used in the MAVA experiment.

4. A number of interesting theories are frequently entertained about the effects of family positions and constellations, of age differences between sibs and between children and parents, upon personality traits. Such theories can be systematically investigated by determining the relative environmental contributions for strategically chosen groups of *S*s in repeated MAVA experiments.

5. The importance of a general "family atmosphere," e.g., in the genesis of neurosis, can be checked through the comparison of within-family and between-family environmental variance on the MAVA applied to measures of neuroticism factors.

6. A variety of theories needs to be tested in regard to the causal directions in discovered correlations of heredity and environment. Does a positive correlation, for intelligence, mean that brighter children create a richer environment or that for social reasons they tend to be born into one? This can be answered partly by the comparison of $r_{be.bh}$ and $r_{we.wh}$. Does the large negative correlation of heredity and environment of the ergic regression factor, UI 23, (Eysenck's "neuroticism factor") require a theory that the neurosis-prone avoid the stresses that might precipitate neurosis? The general finding that hereditary and environmental influences tend to be negatively correlated with the more substantially inherited traits has already been the source of a theory (Cattell, 1957) of "cultural coercion to the bio-social norm."

7. A variety of theories in cultural anthropology and social psychology can now be examined through a comparison of the MAVA experimental results *in different countries*. For example, a democratic organization of society, providing opportunity for all, should raise the importance of heredity relative to environment in determining, for example, performance on intelligence tests.

8. The model also permits examination of theories regarding the effect of one trait upon another. A constitutional endowment of high schizothymia (*A* factor), for example, is to be considered part of the environment for another trait in the same individual, e.g., high intelligence. A comparison of MAVA findings on intelligence ratios for two types of *S*s—those preselected to be high and low, respectively, in *A* factor—can test and shape initial theories.

9. For the personality psychologist the most important theories of all are those concerning the hereditary or environmental origins of particular personality factors. For example, the finding that the *H* primary in the second-order extraversion factor (Cattell & Scheier, 1960) is largely hereditarily determined whereas *F*, surgency-desurgency, is equally a complete product of environmental learning situations, throws considerable new light on extraversion. Also, at one stroke, it cuts the theories to be investigated regarding *H* to largely physiological ones, and those regarding surgency to various possibilities in learning theory. A lot of theoretical wild goose chases are avoided. Equally theoretically provocative is the preliminary finding that Comention, UI 20, which some have called a superego pattern, is actually heavily hereditarily determined.

The above condensed survey may suffice to show that the role of MAVA findings in both forming and testing theoretical developments is considerable, especially in view of the prolonged dearth of evidence in this field and the previous impossibility of formulating results in terms of within- and between-family hereditary and environmental effects, and the correlations among them.

REFERENCES

BURKS, B. S. Nature and nurture: Their influence upon intelligence. Yearb. Nat. Soc. Stud. Educ., 1928, Part 7, 219–316.

BURT, C. L., & HOWARD, M. The relative influence of heredity and environment on assessments of intelligence. *Brit. J. statist. Psychol.*, 1957, **10**, 99–104.

CATTELL, R. B. *The description and measurement of personality.* New York: World Book, 1946.

CATTELL, R. B. Research designs in psychological genetics with special reference to the multiple variance analysis method. *Amer. J. hum. Genet.*, 1953, **5**, 76–93.

CATTELL, R. B. *Personality and motivation structure and measurement.* New York: World Book, 1957.

CATTELL, R. B., BLEWETT, D. B., & BELOFF, J. R. The inheritance of personality: A multiple variance analysis of approximate nature-nurture ratios for primary personality factors in Q-data. *Amer. J. hum. Genet.*, 1955, **7**, 122–146.

CATTELL, R. B., KRISTY, N. F., & STICE, G. F. A first approximation to nature-nurture ratios for eleven primary personality factors in objective tests. *J. abnorm. soc. Psychol.*, 1957, **54**, 143–159.

CATTELL, R. B., & SCHEIER, I. H. *The meaning and measurement of neuroticism and anxiety.* New York: Ronald, 1960.

DANIELS, H. E. The estimation of components of variance. *J. Roy. Statist. Soc.*, 1939, **6**, 186–197.

FISHER, R. A. *The genetical theory of natural selection.* Oxford: Clarendon, 1930.

FISKE, D. W., & RICE, L. Intra-individual response variability. *Psychol. Bull.*, 1955, **52**, 217–250.

HALDANE, J. B. S. *The causes of evolution.* London: Longmans, Green, 1932.

HESS, E. H. The relationship between imprinting and motivation. In, *Nebraska symposium on motivation.* Lincoln: Univer. Nebraska Press, 1959. Pp. 44–77.

KEMPTHORNE, O. *An introduction to genetic statistics.* New York: Wiley, 1957.

MATHER, K. *Biometrical genetics: The study of continuous variation.* New York: Dover, 1949.

MORTON, N. E., & SCHULL, W. F. Studies on consanguinity and heritability. Laboratory of Atomic Bomb Casuality Commission, NAS—NRC, 1953, August 10.

NELDER, J. A. Statistical models in biometrical genetics. *Heredity*, 1953, **7**, 111–119.

THOULESS, R. H. Test unreliability and function fluctuation. *Brit. J. Psychol.*, 1936, **26**, 325–343.

WELCH, W. B. On linear combination of several variances. *J. Amer. Statist. Ass.*, 1956, **51**, 1–9.

WRIGHT, S. Evolution in Mendelian populations. *Genetics*, 1931, **16**, 97–159.

(Received June 15, 1959)

Psychological Review
1960, Vol. 67, No. 6, 373–379

MOTIVATION AND CUE UTILIZATION IN INTENTIONAL AND INCIDENTAL LEARNING[1]

DONALD H. KAUSLER[2] AND E. PHILIP TRAPP

University of Arkansas

In a recent theoretical article, Easterbrook (1959) has attempted to relate level of motivation to the utilization of irrelevant cues in a wide range of behavioral situations. One of these situations pertains to the familiar incidental learning paradigm in which subject (*S*) is simultaneously exposed to relevant and irrelevant cues. Since this procedure provides no orienting instructions to ensure exposure to the irrelevant cues, cognizance of irrelevant cues would appear to be dependent upon the range or span of perception operating during performance on the intentional task (relevant cues). In general, as the range of perception extends beyond the central relevant cues, more peripheral cues are utilized, thus permitting more irrelevant or incidental cue learning; conversely, as the perceptual range constricts, fewer peripheral cues are utilized and incidental cue learning is reduced. According to Easterbrook, drive increments produce a restriction in range of cue use. This restriction, he clearly points out, may facilitate or impair the learning of relevant cues, depending upon the complexity of these cues, but will always impair the learning of peripheral irrelevant cues. As noted by Easterbrook, a similar position has been taken by Bruner, Matter, and Papanek (1955) relating what they call "breadth" of learning to both overmotivation and overlearning.

The present article is concerned further with motivation as a variable in incidental learning. More specifically, it is intended to be an evaluation and an extension of the position taken by Easterbrook that increasing drive level during intentional learning reduces the efficiency of incidental learning through the resulting decrease in irrelevant cue utilization.

However, before proceeding into the crux of the problem, some comments will be made on methodological procedures in incidental-learning experiments which should help serve to define the scope and limitations of the proposed extension.

METHODOLOGIES IN INCIDENTAL LEARNING EXPERIMENTS

In experimental studies investigating the principles of learning as they apply to incidental versus intentional learning, two basic methodologies have evolved. The first methodology requires a group of *S*s to learn a specified task incidentally and, usually, a second group of *S*s to learn the same task intentionally. That is, the intentional group learns the task with specific instructions to learn; the incidental group, on the other hand, is directed toward the task by some orienting instructions but is not instructed to learn the task. The important and critical condition in this methodology is that *S*s in the incidental group are given orienting instructions that assure perception of the learning material without an overt set to learn. Learning of the incidental material is then assumed to indicate the arousal of a covert or generalized set to learn. This methodology has been the pre-

---

[1] This article was supported in part by a grant from the National Science Foundation.

[2] Now at St. Louis University.

dominant one employed in the large number of recent investigations of incidental learning (e.g., Brown, 1954; Postman & Adams, 1956a, 1956b, 1957, 1958; Postman, Adams, & Bohm, 1956; Postman, Adams, & Phillips, 1955; Postman & Phillips, 1954; Saltzman, 1953, 1956, 1957). Studies in this methodological framework typically manipulate systematically either task, procedural, or orienting variables in an attempt to clarify the nature of the processes involved in incidental learning. Rarely, however, is motivation systematically manipulated within this framework. An important exception is a study by Spielberger, Goodstein, and Dahlstrom (1958) in which level of anxiety was investigated as a variable in incidental learning. In this study, however, only an incidental group was used without a corresponding group given instructions to learn.

In the second methodology, each group of *S*s employed in the experiment learns both an intentional task with instructions to learn and a different incidental task without instructions to learn. In other words, in this design, *S* serves as his own control for both intentional and incidental learning. This methodology lends itself more readily to investigating the effects of motivational variables on both intentional and incidental learning than does the first methodology. It is apparent that the restricting influence of motivation formulated by Easterbrook is applicable mainly to this methodology. In fact, four of the five studies cited by Easterbrook (Bahrick, 1954; Kohn, 1954; Silverman, 1954; Silverman & Blitz, 1956) as providing evidence for the effects of motivational concentration on irrelevant cue utilization in human learning are within this framework. The discussion that follows will therefore be restricted primarily to this methodology.

THE NATURE OF MOTIVATION AS A VARIABLE IN CUE UTILIZATION

In his treatment of motivation as a variable in cue utilization, Easterbrook (1959) is referring specifically to emotionally based drives. He states:

> it is an empirically derived generalization that, when the direction of behavior is constant, increase in drive is associated with a reduction in the range of cue use. This paper is devoted to the expansion and exploration of this generalization (pp. 183–184).

He further states:

> The term drive . . . refers to a dimension of emotional arousal or general covert excitement, the innate response to a state of biological deprivation or noxious stimulation, which underlies or occurs simultaneously with overt action and affects its strength and course. This emotional arousal is greater in neurotic than in normal subjects, greater than usual in subjects under stress or threat or in frustration, and in general greater in animals that have been "motivated" by any of the usual deprivations, noxious stimulations, or other incentives than it is in unmotivated or resting animals of the same species (p. 184).

From these statements it seems apparent that Easterbrook is attempting to relate *generalized drive* (*D*), as conceptualized by the Iowa group, to cue utilization.

One of the studies cited by Easterbrook as evidence for motivational restriction, in human learning, however, clearly does not attempt to manipulate generalized *D*. This is the study by Bahrick (1954) in which motivation is heightened by offering financial rewards as an incentive in the high motivational condition. Moreover, in relating motivation to perceptual-motor performance, Easterbrook considers in detail a study by Bahrick, Fitts, and Rankin (1952) in which motivation was again manipulated by financial incentives. This tendency on Easterbrook's part to equate motivational states aroused by specific sets with generalized *D* seems open to serious criticism. It is on this point that the thinking of the present writers begins to differ.

That the nature of the motivation operating during performance on the intentional task may be an important determiner of irrelevant cue utilization is evident from a recent study by Kausler, Trapp, and Brewer (1959) in which they duplicated the intentional and incidental tasks used by Bahrick (1954) with groups motivated by a generalized *D*, anxiety. In the Bahrick situation, motivation was clearly manipulated by an incentive-oriented set. He found that a high incentive group learned a serial list of geometric forms (relevant cues) more efficiently than a low incentive group. However, the high incentive group displayed significantly less incidental learning (irrelevant cues) than the low incentive group in the learning of colors filling in the geometric forms. These results are interpreted by Easterbrook as indicating a restriction or funneling of cue utilization in the high motivational group. The study of Kausler, Trapp, and Brewer (1959), however, yielded conflicting results. In this study the drive condition was generalized emotional *D* rather than incentive-oriented set. Two separate experiments were performed. In the first, anxiety was defined by score on the Taylor MAS; in the second, anxiety was induced by intense ego involving instructions. In both experiments, as in the Bahrick study, the high drive group did significantly better on the intentional task than the low drive group, but, unlike the Bahrick study, no difference was found between the two groups on the incidental task.

In comparing the Bahrick (1954) and Kausler et al. (1959) studies, the differential feature was in the nature of the motivation employed, incentive-oriented set versus generalized *D*. Farber (1955) has emphasized that nongeneralized motives have directive aspects but do not have dynamogenic properties. He states:

> The supposition that motivational variables have steering (associative) properties as well as energizing and reinforcing (drive) properties does not, of course, imply that the two kinds of properties are identical. Indeed, the virtue of the foregoing analyses of the concept of motivation lies in their separation of these functions. The associative function of a motive or any other variable is identified in terms of its tendency to elicit or facilitate a limited class of responses only. The drive function of a variable is demonstrated if: (a) its presence energizes or intensifies indiscriminately all reaction tendencies existing in a given situation; and/or (b) its elimination or reduction in magnitude is reinforcing, i.e., leads to the increased probability of recurrence, in the same situation, of the responses preceding its modification. Motives, it is presumed, have both functions, and if a variable clearly does not have both, its status as a motive is questionable (pp. 311–312).

Thus, incentive-oriented sets would seem to possess associative properties without drive properties. That is, they elicit particular responses that are goal directed but do not energize other responses that are irrelevant to the goal. In general, sets of this nature are experimentally induced by encouraging efficient performance on the relevant task without intentionally arousing emotional responses. This is typically accomplished by means of rewards or by task-oriented instructions. Generalized *D*, on the other hand, serves an energizing function in that its presence intensifies indiscriminately all reaction tendencies existing in a given situation, both relevant and irrelevant to the central task. In general, *D* is experimentally manipulated by the intentional arousal of emotional responses during the learning process. This is typically accomplished by differentiating between *S*s on some measure of intrinsic anxiety or by inducing anxiety through the use of ego involving instructions, failure, stress, etc.

It is conceivable that in both the Bahrick and Kausler et al. studies increments in motivation, either associative or drive, resulted in a reduction of habit strength for the irrelevant cues through the restriction of perceptual range proposed by Easterbrook. Thus, the high motivational groups acquired less irrelevant habit strength than the low motivational group. In the Bahrick study this difference in habit strength is manifested because the difference in motivation between the two groups is relevant only to the intentional task and, therefore, does not enter into the multiplicative relationship (cf. Spence, 1958a, 1958b) between irrelevant habit strength and drive level. In the Kausler et al. study the generalized *D* would be expected to intensify the relatively weaker habit strength in the high drive group, thus bringing its response potential to a point comparable to that in the low drive group where habit strength is stronger but drive level is weaker. Hence, the groups would not differ in incidental learning.

POSITIONING OF RELEVANT AND IRRELEVANT CUES

The results of the Kausler et al. study also conflict with the studies by Silverman (1954) and Silverman and Blitz (1956) cited by Easterbrook as yielding supporting evidence for the restrictive influence of high motivation. In all three studies motivation was manipulated by an emotionally based generalized *D*. In the Silverman study anxiety was induced by threat of shock, and in the Kausler et al. study it was induced by ego involving instructions. In both the Silverman and Blitz and Kausler et al. studies intrinsic anxiety was employed by selecting extreme scorers on the MAS.

In both Silverman studies an increase in anxiety level was followed by the expected decrease in incidental learning in constrast to the no difference reported in the Kausler et al. study. Although the drive state does not differ in the three studies, there is another important differential between them. This is in terms of the relationship of the irrelevant cues to the relevant cues within the perceptual field. In the Kausler et al. study, as in the Bahrick study, the irrelevant cues are embodied within the relevant cues (colors within geometric forms), thus assuring close proximity, both spatially and temporally, between relevant and irrelevant cues. Consequently, even though anxious *S*s may have a restricted perceptual range, the irrelevant cues may well fall within that range. In contrast, the relevant and irrelevant cues in the Silverman studies are widely spaced within the perceptual field. In fact, in the Silverman (1954) study, the relevant cues are perceived by a different sense modality (vision) than are the irrelevant cues (audition). In the Silverman and Blitz (1956) study, the irrelevant cues (two-digit numbers) occupied a different section of a memory drum opening than did the relevant cues (nonsense syllables), and were separated from the relevant cues by six centimeters. It is very possible, therefore, that in the Kausler et al. study the differential perception between high and low anxious groups was not as pronounced as in the Silverman studies.

It is conceivable then that the amount of incidental learning under high and low motivation is affected differentially by position of the irrelevant cues within the perceptual field. That is, there is an interaction effect between level of motivation and cue positioning. In general, as the irrelevant cues are moved toward the periphery of the perceptual field, incidental learning will decrease, but the reduction will be greater for the high motivational levels than for low. Considerable systematic investigation is needed to establish the validity of this proposed hypothesis.

TASK AND PROCEDURAL VARIABLES

Other important variables to take into account in evaluating the effects of increased motivation on intentional and incidental learning are variables relating to the nature of the intentional task and method of presenting the intentional task. Particularly important variables would seem to be variables relating to the difficulty of the intentional task, such as, length, intratask competition, rate of presentation, etc. A number of studies employing the first methodology discussed earlier have systematically investigated these variables and demonstrated their importance in incidental learning studies (e.g., Neimark & Saltzman, 1953; Postman & Adams, 1957; Postman, Adams, & Phillips, 1955).

Differences in the tasks used, both intentional and incidental, make it more arduous to compare the results of studies with the simultaneous relevant-irrelevant cue methodology. This is especially true when motivation is varied qualitatively from study to study. Kausler and Trapp (1959) have suggested that studies making cross-drive comparisons should use, whenever possible, the same task in order to control for task effects.

Some evidence that the efficiency of incidental learning under different levels of motivation may be related to task difficulty is indicated in the study of Spielberger, Goodstein, and Dahlstrom (1958). Using the first methodology but without an intentional group, they found that the relationship between anxiety level (as defined by score on the MAS) and incidental learning (learning the designs on the Bender-Gestalt test) was a function of the difficulty of the learning material. On the easy designs the high anxious *S*s performed at a higher level than the low anxious *S*s; on the difficult designs the low anxious *S*s performed at a higher level than the high anxious *S*s.

The present writers believe that task difficulty variables operate in the intentional-incidental learning situation by altering the degree of concentration required of *S*s. Rate of presentation, for example, would be important in that the faster the rate, the greater the degree of concentration demanded. It is here suggested that intentional task variables which alter the degree of concentration required of *S* will affect further the range of attention for simultaneously present irrelevant ones. However, the net effect on the learning of irrelevant cue habits should also be a function of the nature of the motivation present (incentive-oriented versus generalized *D*), the level of motivation, and the positioning of irrelevant cues within the perceptual field. Thus studies relating degree of motivation to irrelevant cue utilization should also take into account the nature of the drive, the nature of the intentional task, and the positioning of the irrelevant cues.

For example, on easy tasks which require little concentration, high generalized drive *S*s should display more incidental learning than low drive *S*s (as in the Spielberger et al., 1958, study). The perceptual range should be fairly broad under these conditions and should incorporate irrelevant cues within close proximity to the relevant cues. Since generalized *D* reacts multiplicatively with irrelevant habit strength, the reaction tendency for high drive *S*s should be greater than for low drive *S*s. On difficult tasks, however, the perceptual range for irrelevant cues should be much narrower. Consequently, when combined with the further restricting effect of high drive, the attention for these cues and subsequent habit strength should be considerably lower for high drive *S*s than for low drive *S*s. Under these conditions, low drive *S*s should begin to display superior incidental learning.

With incentive-oriented motivation, however, low motivation might be expected to result in superior incidental learning regardless of intentional task difficulty. This is likely because the motivation involved does not enter into a multiplicative relationship with habit strength. Nevertheless, the difference in incidental learning proficiency between high and low incentive groups could be expected to diminish as the intentional task decreases in difficulty. That is, the highly motivated *S*s would need to concentrate less on easy tasks, thus reducing the funneling effect on their attention to irrelevant cues; conversely, on difficult tasks requiring intense concentration, the funneling effect of high incentive motivation should be most pronounced.

It is obvious that considerable systematic research is needed, manipulating task variables along with motivational and positional variables, before the overall effects of level of motivation on irrelevant cue utilization can be determined. It does seem clear, however, that the phenomena Easterbrook seeks to embrace are too complex to be unified under one general principle, appealing as the thought may be. Even at this stage, the empirical data demand some extensions to his formulations. The elaborations presented here do reconcile existing conflicting findings. How well they can handle derivable hypotheses, like those suggested in the foregoing, remains to be seen. Studies to that effect are currently in process in our laboratories.

SUMMARY

Easterbrook has proposed that increasing motivation (specifically defined as emotional *D*) results in a funneling or focusing of attention to relevant cues when *S*s are simultaneously exposed to relevant and irrelevant cues. The present article suggests that the net effects of increasing motivation on cue utilization during incidental learning are functions of additional variables as well. Of particular importance seem to be the nature of the motivation (generalized *D* versus incentive-oriented set), position of the irrelevant cues relative to the relevant cues (central versus peripheral), and difficulty of the intentional task. This extension of Easterbrook's framework brings on a rapprochement with existing research findings and opens the way to more systematic investigations into an admittedly complex area.

REFERENCES

BAHRICK, H. P. Incidental learning under two incentive conditions. *J. exp. Psychol.*, 1954, **47**, 170–172.

BAHRICK, H. P., FITTS, P. M., & RANKIN, R. E. Effects of incentives upon reactions to peripheral stimuli. *J. exp. Psychol.*, 1952, **44**, 400–406.

BROWN, G. H. Factors influencing incidental learning. *J. exp. Psychol.*, 1954, **47**, 163–169.

BRUNER, J. S., MATTER, JEAN, & PAPANEK, MIRIAM. Breadth of learning as a function of drive level and mechanization. *Psychol. Rev.*, 1955, **62**, 1–10.

EASTERBROOK, J. A. The effect of emotion on cue utilization and the organization of behavior. *Psychol. Rev.*, 1959, **66**, 183–201.

FARBER, I. E. The role of motivation in verbal learning and performance. *Psychol. Bull.*, 1955, **52**, 311–327.

KAUSLER, D. H., & TRAPP, E. P. Methodological considerations in the construct validation of drive-oriented scales. *Psychol. Bull.*, 1959, **56**, 152–157.

KAUSLER, D. H., TRAPP, E. P., & BREWER, C. L. Intentional and incidental learning under high and low emotional drive levels. *J. exp. Psychol.*, 1959, **58**, 452–455.

KOHN, H. Effects of variations of intensity of experimentally induced stress situations upon certain aspects of perception and performance. *J. genet. Psychol.*, 1954, **85**, 289–304.

NEIMARK, EDITH, & SALTZMAN, I. J. Intentional and incidental learning with different rates of stimulus presentation. *Amer. J. Psychol.*, 1953, **66**, 618–621.

POSTMAN, L., & ADAMS, PAULINE, A. Studies in incidental learning: III. Interserial interference. *J. exp. Psychol.*, 1956, **51**, 323–328. (a)

POSTMAN, L., & ADAMS, PAULINE A. Studies in incidental learning: IV. The interaction of orienting tasks and stimulus materials. *J. exp. Psychol.*, 1956, **51**, 329–333. (b)

POSTMAN, L., & ADAMS, PAULINE A. Studies in incidental learning: VI. Intraserial interference. *J. exp. Psychol.*, 1957, **54**, 153–167.

POSTMAN, L., & ADAMS, PAULINE A. Studies in incidental learning: VII. Effects of frequency of exercise and length of list. *J. exp. Psychol.*, 1958, **56**, 86–93.

POSTMAN, L., ADAMS, PAULINE A., & BOHM, A. M. Studies in incidental learning: V. Recall for order and associative clustering. *J. exp. Psychol.*, 1956, **51**, 334–342.

POSTMAN, L., ADAMS, PAULINE A., & PHILLIPS, LAURA W. Studies in incidental learning: II. The effects of association value and the method of testing. *J. exp. Psychol.*, 1955, **49**, 1–10.

POSTMAN, L., & PHILLIPS, LAURA W. Studies in incidental learning: I. The effects of crowding and isolation. *J. exp. Psychol.*, 1954, **48**, 48–56.

SALTZMAN, I. J. The orienting task in incidental and intentional learning. *Amer. J. Psychol.*, 1953, **66**, 593–597.

SALTZMAN, I. J. Comparisons of incidental and intentional learning with different orienting tasks. *Amer. J. Psychol.*, 1956, **69**, 274–277.

SALTZMAN, I. J. Incidental and intentional memory for lifted-weights. *Amer. J. Psychol.*, 1957, **70**, 253–257.

SILVERMAN, R. E. Anxiety and the mode of response. *J. abnorm. soc. Psychol.*, 1954, **49**, 538–542.

SILVERMAN, R. E., & BLITZ, B. Learning and two kinds of anxiety. *J. abnorm. soc. Psychol.*, 1956, **52**, 301–303.

SPENCE, K. W. A theory of emotionally based drive (D) and its relation to performance in simple learning situations. *Amer. Psychologist*, 1958, **13**, 131–141. (a)

SPENCE, K. W. Behavior theory and selective learning. In M. R. Jones (Ed.), *Nebraska symposium on motivation*. Lincoln: Univer. Nebraska Press, 1958. Pp. 73–107. (b)

SPIELBERGER, C. D., GOODSTEIN, L. D., & DAHLSTROM, W. G. Complex incidental learning as a function of anxiety and task difficulty. *J. exp. Psychol.*, 1958, **56**, 58–61.

(Received September 29, 1959)

Psychological Review
1960, Vol. 67, No. 6, 380–400

BLIND VARIATION AND SELECTIVE RETENTION IN CREATIVE THOUGHT AS IN OTHER KNOWLEDGE PROCESSES[1]

DONALD T. CAMPBELL

Northwestern University

This paper proposes to examine creative thought within the framework of a comparative psychology of knowledge processes, and in particular with regard to one theme recurrent in most knowledge processes. This theme may be expressed as follows:

1. A blind-variation-and-selective-retention process is fundamental to all inductive achievements, to all genuine increases in knowledge, to all increases in fit of system to environment.
2. The many processes which shortcut a more full blind-variation-and-selective-retention process are in themselves inductive achievements, containing wisdom about the environment achieved originally by blind variation and selective retention.
3. In addition, such shortcut processes contain in their own operation a blind-variation-and-selective-retention process at some level, substituting for overt locomotor exploration or the life-and-death winnowing of organic evolution.

Between a modern experimental physicist and some virus-type ancestor there has been a tremendous gain in knowledge[2] about the environment. In bulk, this has represented cumulated inductive achievements, stage by stage expansions of knowledge beyond what could have been deductively derived from what had been previously known. It has represented repeated "breakouts" from the limits of available wisdom, for if such expansions had represented only wise anticipations, they would have been exploiting full or partial knowledge already achieved. Instead, real gains must have been the products of explorations going beyond the limits of foresight or prescience, and in this sense blind. In the instances of such real gains, the successful explorations were in origin as blind as those which failed. The difference between the successful and unsuccessful was due to the nature of the environment encountered, representing discovered wisdom about that environment.

---

[1] A partially overlapping version of this paper was presented at the Inter-Disciplinary Conference on Self-Organizing Systems, sponsored by the Office of Naval Research and the Armour Research Foundation of the Illinois Institute of Technology, Chicago, May 5–6, 1959. The proceedings of the conference are to be published by Pergamon Press under the title *Self-Organizing Systems*. The author is indebted to Carl P. Duncan for contributing to the development of many of the points involved.

[2] This extended usage of "knowledge" is a part of an effort to put "the problem of knowledge" into a behavioristic framework which takes full cognizance of man's status as a biological product of an evolutionary development from a highly limited background, with no "direct" dispensations of knowledge being added at any point in the family tree. The bibliographical citation of the several sources converging on this approach to the problem of knowledge, and the discussion of its relation to traditional philosophical issues and to the strategy of science are presented elsewhere (Campbell, 1959). Suffice it to say here that the position limits one to "an epistemology of the other one." The "primitives" of knowledge can not be sought in "raw feels" or in "phenomenal givens," or in any "incorrigible" elements. While man's conscious knowledge processes are recognized as more complex and subtle than those of lower organisms, they are not taken as more fundamental or primitive. In this perspective, any process providing a stored program for organismic adaptation in external environments is included as a knowledge process, and any gain in the adequacy of such a program is regarded as a gain in knowledge. If the reader prefers, he can understand the paper adequately regarding the term "knowledge" as metaphorical when applied to the lower levels in the developmental hierarchy. But since the problem of knowledge has resisted any generally accepted solution when defined in terms of the conscious contents of the philosopher himself, little seems lost and possibly something gained by thus extending the range of processes considered.

The general model for such inductive gains is that underlying both trial-and-error problem solving and natural selection in evolution, the analogy between which has been noted by several persons (e.g., Ashby, 1952; Baldwin, 1900; Pringle, 1951). Three conditions are necessary: a mechanism for introducing variation, a consistent selection process, and a mechanism for preserving and reproducing the selected variations. In what follows we shall look for these three ingredients at a variety of levels. But first a comment on the use of the word "blind" rather than the more usual "random." It seems likely that Ashby (1952) unnecessarily limited the generality of his mechanism in Homeostat by an effort to fully represent all of the modern connotations of random. Equiprobability is not needed, and is definitely lacking in the mutations which lay the variation base for organic evolution. Statistical independence between one variation and the next, while frequently desirable, can also be spared: in particular, for the generalizations essayed here, certain processes involving systematic sweep scanning are recognized as blind, insofar as variations are produced without prior knowledge of which ones, if any, will furnish a selectworthy encounter. An essential connotation of blind is that the variations emitted be independent of the environmental conditions of the occasion of their occurrence. A second important connotation is that the occurrence of trials individually be uncorrelated with the solution, in that specific correct trials are no more likely to occur at any one point in a series of trials than another, nor than specific incorrect trials. (Insofar as observation shows this not to be so, the system is making use of already achieved knowledge, perhaps of a general sort. The prepotent responses of an animal in a new puzzle box toward the apparent openings may thus represent prior general knowledge, transferred from previous learning or inherited as a product of the mutation and selective survival process.) A third essential connotation of blind is rejection of the notion that a variation subsequent to an incorrect trial is a "correction" of the previous trial or makes use of the direction of error of the previous one. (Insofar as mechanisms do seem to operate in this fashion, there must be operating a substitute process carrying on the blind search at another level, feedback circuits selecting "partially" adequate variations, providing information to the effect that "you're getting warm," etc.)

REVIEW OF THE THEME IN LOWER KNOWLEDGE PROCESSES

In this perspective, the epistemologically most fundamental knowledge processes are embodied in those several inventions making possible organic evolution. At the already advanced level of cellular life, this is a "learning" on the part of the species by the blind variation and selective survival of mutant individuals. In terms of the three requirements, variation is provided by the mutations, selection by the somewhat consistent or "knowable" vagaries of the environment, and preservation and duplication by the complex and rigid order of chromosome mitosis. Bisexuality, heterozygosity, and meiotic cell division represent a secondary invention increasing the efficiency of the process through increasing the range of variation and the rate of readjustment to novel environments. The selection and preservation processes remain the same. The ubiquity of bisexuality, its several independent inventions, and the multifarious elaboration of the theme, all speak to its tremendous usefulness.

The higher evolutionary developments shift a part of the locus of adaptation away from a trial and error of whole organisms or gene pools, over to processes occurring within the single organism. Such processes are numerous, each being not only a device for obtaining knowledge, but also representing general wisdom about environmental contingencies already achieved through organic evolution, making possible more efficient achievement of detailed local knowledge. One of the most primitive of these is exploratory locomotion, described in the protozoa by Jennings (1906) and accepted as a model for Homeostat by Ashby (1952). Forward locomotion persists until blocked, at which point direction of locomotion is varied blindly until unblocked forward locomotion is again possible. The external physical environment is the selection agency, the preservation of discovery is embodied in the perseveration of the unblocked forward movement. At this level, the species has "discovered" that the environment is discontinuous, consisting of penetrable regions and impenetrable ones, and that impenetrability is to some extent a stable characteristic—it has discovered that when blocked it is a better strategy to try to go around than to wait until one can move through.

Insofar as individual organisms without distance receptors (such as paramecia and earthworms) can learn through contiguity, the species has already achieved the more general knowledge that there is some event-contingency stability in the environment. That is, in the degree to which individual learning is useful, there has been the species-level discovery of slower transformation processes on the part of relevant segments of the environment than of the organism. In addition, whereas the ultimate selection is life or death in encounters with the external environment, by the evolutionary stage at which learning is possible, much of this once-external criterion has been internalized. Crude environmental contingencies with low selection ratios are now represented as pleasures or pains, or as reinforcers more generally. The selection becomes much more sharp, but the contact with the environmental realities less direct.

The presence of a fundamental trial-and-error process in individual learning needs no elaboration or defense. Suffice it to say that recognition of such a process is found in all learning theories which make any pretense of completeness, including at least three of Gestalt inspiration (Campbell, 1956a). While higher vertebrate (and higher cephalopod) learning makes far more use of the short circuiting of overt trial and error by vision than is allowed for by the usual learning theory (Campbell, 1956b), for convenience here the multiplication of levels will be avoided by treating trial-and-error learning as a single process level.

The next and most striking class of discoveries are those centering around echo-location and vision. Woodworth (1921) has emphasized the achievement of a percept from the ingredients of sensation through a series of "trial-and-error perceptions." Thurstone (1924) has interpreted perception as a trial and error of potential locomotions placed in a hierarchy of trial-and-error processes including both overt trial and error and ideational trial and error, in a book containing many anticipations of cybernetic concepts. Pumphrey (1950) interprets the primitive sense receptor of the fishes called the "lateral-line organ" as a crude echo-location device, making use of the reflected pulses of the fish's own swimming. Griffin (1958) has documented in detail the use by bats and cave birds of sonic and supersonic vocalizations selectively reflected by obstacles of the environment. Kellogg (1958) has made a similar case for the porpoise. Here is a powerful substitute for blind locomotor exploration. (See Simon, 1957, p. 264, for an estimate of such gains.) In echo location a wave pulse is emitted blindly in all directions. The obstacles of the environment selectively reflect the pulse from certain of these directions, and thus provide a feedback which is substitutable for that which would have been received had the animal locomoted in those directions. Radar guidance systems employ an analogous substitution of a blindly scanning electromagnetic wave pulse, in economical substitution for a blind scanning of the same environment of potential locomotions by full ship or projectile movements.

Visual perception seems interpretable as a substitute search process of similar order (Campbell, 1956b). The full analogy is weakened by the absence of an emitting process on the part of the organism. Instead, advantage is taken of diffuse electromagnetic waves made available from external sources. Consider first a pseudoeye consisting of but a single photoreceptor cell. (Such a device has been distributed for use by the blind in which a photocell output is transformed into a sound of variable pitch.) With such a device, blind scanning as in a radar system is essential. Brightness contours can be located and fixated by continual crossing, as in the "hunting" process in a mechanical servosystem, or as in the vocal pitch control in which a steady note is "held" only by a continuous search oscillation (Deutsch & Clarkson, 1959). To conceive of such an "eye" as a blind searching device substituting for a more costly blind locomotion in the explored directions is not difficult. The eyes of insects and vertebrates and the higher cephalopods differ from such a device by having multiple photocells, making possible selective reflection from objects in multiple directions at once. Each receptor cell can be conceived of as exploring the possibilities of locomotion in a given direction, the retina collectively thus exploring the possibilities of locomotion in a wide segment of potential directions for locomotion. Except as the eye is aimed by other sources of knowledge, these possibilities have been made "blindly" available without prescience or insight. For the "blindness" of an eyeless animal there has been substituted a process so efficient that we use it naively as a model for direct, unmediated knowing. But the process is still one of blind search and selective retention, in the sense employed in this paper.

Vision is a very complex and marvelous mechanism, and the brief presentation here does not do justice even to the random search components involved. Hebb (1949) has well documented the active search of eye movements, correcting the model of the inactive fixed-focus eye which is implicit in both Gestalt psychology and conditioning theory. Riggs (Riggs, Armington, & Ratliff, 1954) and Ditchburn (1955) have documented the essential role of the continuous low amplitude scanning provided by "physiological nystagmus" or "fixation tremor." Platt (1958) has provided a brilliant analysis of the role of a blind "rubbing" process, his "lens-grinding" model, for the achievement of visual acuity and spatial representation in a visual system containing unaddressed elements. These and other considerations convince the present writer that although vision represents the strongest challenge to the generality a blind-variation-and-selective-retention aspect to all knowledge processes, it is not in fact an exception. These brief comments have not fully justified this conclusion, however.

Taking these echo location and visual exploratory processes collectively, several general aspects can be noted: all exploit a specific and limited coincidence, i.e., that objects impenetrable by organismic locomotion also are opaque to, or reflect, certain wave forms in the acoustical frequencies and in the bands of electromagnetic waves of the visual and radar spectra. It is this coincidence, unpredictable upon the basis of the prior knowledge available to the more primitive organisms, which makes possible such marvelously efficient shortcuts. Thus while phenomenologically vision is more direct than other knowledge processes, it is seen in this perspective as an indirect, substitute process. As in all substitute knowledge processes, the effectiveness is limited by the accuracy of the coding process, i.e., the translation terms between one level and another. Such coding is never exhaustive (Platt, 1956). It always involves abstraction, and along with this some fringe imperfection and proneness to systematic error. It must finally be checked out and corrected by overt locomotion. Its efficacy is limited by the *relevance* of the coding to the more fundamental level of behavior for which it is a substitute. This relevance was itself initially tested out by a blind-variation-and-selective-retention process at the level of organic evolution or early childhood learning. (Species differ in this regard.) The phenomenal directness of vision tempts us to make vision prototypic for knowing at all levels, and leads to that chronic belief in the potential existence of direct and "insightful" mental processes, a belief which it is one purpose of this paper to deny.

CREATIVE THOUGHT

Creative thought provides the next level knowledge process for the present discussion. At this level there is a *substitute* exploration of a *substitute* representation of the environment, the "solution" being selected from the multifarious exploratory thought trials according to a criterion which is in itself *substituting* for an external state of affairs. Insofar as the three substitutions are accurate, the solutions when put into overt locomotion are adaptive, leading to intelligent behavior which lacks overt blind floundering, and is thus a knowledge process. To include this process in the general plan of blind-variation-and-selective-retention, it must be emphasized that insofar as thought achieves innovation, the internal emitting of thought trials one by one is blind, lacking prescience or foresight. The process *as a whole* of course provides "foresight" for the overt level of behavior, once the process has blindly stumbled into a thought trial that "fits" the selection criterion, accompanied by the "something clicked," "Eureka," or "aha-erlebnis" that usually marks the successful termination of the process.

Today, we find the blind-variation-and-selective-retention model most plausibly applied at the levels of organic evolution and trial-and-error learning of animals, and least palatable as a description of creative thinking. Historically, however, the phrase "trial and error" was first used to describe thinking, by Alexander Bain as early as 1855, two years before Darwin's publication of the doctrine of natural selection. Not only for historical interest, but also to further develop the psychology of creativity, the following quotations from him (Bain, 1874) are provided:

> Possessing thus the material of the construction and a clear sense of the fitness or unfitness of each new tentative, the operator proceeds to ply the third requisite of constructiveness—trial and error— . . . to attain the desired result. . . . The number of trials necessary to arrive at a new construction is commonly so great that without something of an affection or fascination for the subject one grows weary of the task. This is the *emotional* condition of originality of mind in any department (p. 593).
>
> In the process of Deduction . . . the same constructive process has often to be introduced. The mind being prepared beforehand with the principles most likely for the purpose . . . incubates in patient thought over the problem, trying and rejecting, until at last the proper elements come together in the view, and fall into their places in a fitting combination (p. 594).
>
> With reference to originality in all departments, whether science, practice, or fine art, there is a point of character that deserves notice. . . . I mean an Active turn, or a profuseness of energy, put forth in trials of all kinds on the chance of making lucky hits . . . Nothing less than a fanaticism of experimentation could have given birth to some of our grandest practical combinations. The great discovery of Daguerre, for example, could not have been regularly worked out by any systematic and orderly research; there was no way but to stumble upon it. . . . The discovery is unaccountable, until we learn that the author . . . got deeply involved in trials and operations far removed from the beaten paths of inquiry (p. 595).

In 1881 Paul Souriau presented a still more preponderant emphasis on the factor of chance as the sole source of true innovation. He asserts again and again that "le principe de l'invention est le hasard." In the main, he presents his argument through illustration and through the elimination of rival hypotheses about the inventive process, including deduction, induction, and "la methode." A positive explanation of the process is hard to find. This sample will illustrate his approach:

> A problem is posed for which we must invent a solution. We know the conditions to be met by the sought idea; but we do not know what series of ideas will lead us there. In other words, we know how the series of our thoughts must end, but not how it should begin. In this case it is evident that there is no way to begin except at random. Our mind takes up the first path that it finds open before it, perceives that it is a false route, retraces its steps and takes another direction. Perhaps it will arrive immediately at the sought idea, perhaps it will arrive very belatedly: it is entirely impossible to know in advance. In these conditions we are reduced to dependence upon chance.
>
> In the case just analysed we supposed that we had to solve a problem already stated for us. But how was the problem-statement itself found? It is said that a question well posed is half answered. If so, then true invention consists in the posing of questions. There is something mechanical, so to speak, in the art of finding solutions. The truly original mind is that which discovers problems. But here again, it does no good to speak of method, since method is the application of already existing discoveries. The discovery of a new problem can therefore only be fortuitous. Thus we see the role of logic diminish and that of chance increase as we approach closer to true invention. Chance is the first principle of invention: it is what has produced method, nourished it, and made it fertile. Method can only analyse the ideas which come to it from elsewhere, drawing out their consequences and exhausting their contents. Left to itself method soon becomes sterile. Methodological minds cannot help having a feeling of disdain for adventurous minds which affirm before proving and believe before knowing. But they must recognize that without such audacity, no progress would be possible. The mind is not able to revise itself upon its own

> foundations. New ideas cannot have prototypes: their appearance can only be attributed to chance (pp. 17–18).

Souriau has not only the notion of chance combinations, but also the concept of their being produced in large numbers which are generally worthless and from which only the rare ones fitting a goal or criterion are selected. These two widely separated quotations illustrate this:

> By a kind of artificial selection, we can in addition substantially perfect our thought and make it more and more logical. Of all of the ideas which present themselves to our mind, we note only those which have some value and can be utilized in reasoning. For every single idea of a judicious and reasonable nature which offers itself to us, what hosts of frivilous, bizarre, and absurd ideas cross our mind. Those persons who, upon considering the marvelous results at which knowledge has arrived, cannot imagine that the human mind could achieve this by a simple fumbling, do not bear in mind the great number of scholars working at the same time on the same problem, and how much time even the smallest discovery costs them. Even genius has need of patience. It is after hours and years of meditation that the sought-after idea presents itself to the inventor. He does not succeed without going astray many times; and if he thinks himself to have succeeded without effort, it is only because the joy of having succeeded has made him forget all the fatigues, all of the false leads, all of the agonies, with which he has paid for his success (p. 43).
>
> . . . If his memory is strong enough to retain all of the amassed details, he evokes them in turn with such rapidity that they seem to appear simultaneously; he groups them by chance in all the possible ways; his ideas, thus shaken up and agitated in his mind, form numerous unstable aggregates which destroy themselves, and finish up by stopping on the most simple and solid combination (pp. 114–115).

The phrase "artificial selection" is reminiscent of Darwin's writings, although Souriau makes no mention of the selective-survival model of evolution, (nor does he cite the ideas of any other person whatsoever).[3] Note how similar the final quotation is to Ashby's (1952) phrasing of the inevitable self-elimination of unstable combinations:

> Just as, in the species, the truism that the dead cannot breed implies that there is a fundamental tendency for the successful to replace the unsuccessful, so in the nervous system does the truism that the unstable tends to destroy itself imply that there is a fundamental tendency for the stable to replace the unstable (p. vi).

Ernst Mach was another great 19th century thinker about thinking who emphasized this model. We today remember him most as a psychologist-physicist-philosopher who contributed to the present day positivistic recognition of the hypothetic character of our constructions of the world and who first made explicit the empirical presumptions involved in the physicist's assumption of an Euclidian space. But when, at the age of 57 in 1895, he was called back to his alma mater the University of Vienna to assume a newly created position of Professor of the History and Theory of Inductive Science, he chose a quite different theme for his inaugural address. His title was "on the part played by accident in invention and discovery." The occasion indicates the importance he gave to the message, and indeed, his paper is a neglected classic in the psychology of knowledge processes. These quotations (Mach, 1896) further reinforce the model of creative thought being presented:

---

[3] Souriau's presentation is in general quite modern in spirit, although associationistic in a way some would find dated, and vigorously deterministic in a way now undermined by subatomic physics, although not necessarily so for the problems of which he treats. He comments wisely on many topics not covered here, including simultaneous independent invention and the Zeitgeist, the social conditions of creativity and invention, the dissonance created by discrepant opinions of others, the congruence of free will and determinism, and both the conflict and interdependency between erudition and innovation. His attacks on both deduction and induction are reminiscent of Peirce's later critiques.

> The disclosure of new provinces of facts before unknown, can only be brought about by accidental circumstances . . . (p. 168).
>
> In such [other] cases it is a psychical accident to which the person owes his discovery—a discovery which is here made "deductively" by means of mental copies of the world, instead of experimentally (p. 171).
>
> After the repeated survey of a field has afforded opportunity for the interposition of advantageous accidents, has rendered all the traits that suit with the word or the dominant thought more vivid, and has gradually relegated to the background all things that are inappropriate, making their future appearance impossible; then from the teeming, swelling host of fancies which a free and high-flown imagination calls forth, suddenly that particular form arises to the light which harmonizes perfectly with the ruling idea, mood, or design. Then it is that that which has resulted slowly as the result of a gradual selection, appears as if it were the outcome of a deliberate act of creation. Thus are to be explained the statements of Newton, Mozart, Richard Wagner, and others, when they say that thoughts, melodies, and harmonies had poured in upon them, and that they had simply retained the right ones (p. 174).

Poincaré (1908, 1913) in his famous essay on mathematical invention presents a point of view which is also judged to be in agreement. He first gives an example in imagery: "One evening, contrary to my custom, I drank black coffee and could not sleep. Ideas rose in crowds; I felt them collide until pairs interlocked, so to speak, making a stable combination" (Poincaré, 1913, p. 387). Poincaré feels that it is rare for this blind permuting process to rise into conscious awareness, and that as a rule only the successful selected alternatives enter consciousness. Because of the relevance of Poincaré's comments; because Hadamard (1945) has cited him in opposition to the accidentalist position while he is read here as favoring the selective-retention version of it; and because of all of the sources cited he would most generally be respected as truly creative (in the field of mathematics) these longish excerpts (Poincaré, 1913) are read into the record:

> It is certain that the combinations which present themselves to the mind in a sort of sudden illumination, after an unconscious working somewhat prolonged, are generally useful and fertile combinations, which seem the result of a first impression. Does it follow that the subliminal self, having divined by a delicate intuition that these combinations would be useful, has formed only these, or has it rather formed many others which were lacking in interest and have remained unconscious?
>
> In this . . . way of looking at it, all the combinations would be formed in consequence of the automatism of the subliminal self, but only the interesting ones would break into the domain of consciousness. And this is still very mysterious. What is the cause that, among the thousand products of our unconscious activity, some are called to pass the threshold, while others remain below? Is it a simple chance which confers this privilege? Evidently not; among all the stimuli of our senses, for example, only the most intense fix our attention, unless it has been drawn to them by other causes. More generally the privileged unconscious phenomena, those susceptible of becoming conscious, are those which, directly or indirectly, affect most profoundly our emotional sensibility (p. 391).
>
> . . . we reach the following conclusion: The useful combinations are precisely the most beautiful, I mean those best able to charm this special sensibility that all mathematicians know, but of which the profane are so ignorant as often to be tempted to smile at it.
>
> What happens then? Among the great numbers of combinations blindly formed by the subliminal self, almost all are without interest and without utility; but just for that reason they are also without effect upon the esthetic sensibility. Consciousness will never know them; only certain ones are harmonious, and, consequently, at once useful and beautiful. They will be capable of

touching this special sensibility of the geometer of which I have just spoken, and which, once aroused, will call our attention to them, and thus give them occasion to become conscious.

This is only a hypothesis, and yet here is an observation which may confirm it: when a sudden illumination seizes upon the mind of the mathematician, it usually happens that it does not deceive him, but it also sometimes happens, as I have said, that it does not stand the test of verification; well, we almost always notice that this false idea, had it been true, would have gratified our natural feeling for mathematical elegance.

Thus it is this special esthetic sensibility which plays the role of the delicate sieve of which I spoke, and that sufficiently explains why the one lacking it will never be a real creator.

Yet all the difficulties have not disappeared. The conscious self is narrowly limited, and as for the subliminal self we know not its limitations, and this is why we are not too reluctant in supposing that it has been able in a short time to make more different combinations than the whole life of a conscious being could encompass. Yet these limitations exist. Is it likely that it is able to form all the possible combinations, whose number would frighten the imagination? Nevertheless that would seem necessary, because if it produces only a small part of these combinations, and if it makes them at random, there would be small chance that the *good*, the one we should choose, would be found among them.

Perhaps we ought to seek the explanation in that preliminary period of conscious work which always precedes all fruitful unconscious labor. Permit me a rough comparison. Figure the future elements of our combinations as something like the hooked atoms of Epicurus. During the complete repose of the mind, these atoms are motionless, they are, so to speak, hooked to the wall; so this complete rest may be indefinitely prolonged without the atoms meeting, and consequently without any combination between them.

On the other hand, during a period of apparent rest and unconscious work, certain of them are detached from the wall and put in motion. They flash in every direction through the space (I was about say the room) where they are enclosed, as would, for example, a swarm of gnats or, if you prefer a more learned comparison, like the molecules of gas in the kinematic theory of gases. Then their mutual impacts may produce new combinations.

What is the role of the preliminary conscious work? It is evidently to mobilize certain of these atoms, to unhook them from the wall and put them in swing. We think we have done no good, because we have moved these elements a thousand different ways in seeking to assemble them, and have found no satisfactory aggregate. But, after this shaking up imposed upon them by our will, these atoms do not return to their primitive rest. They freely continue their dance.

Now, our will did not choose them at random; it pursued a perfectly determined aim. The mobilized atoms are therefore not any atoms whatsoever; they are those from which we might reasonably expect the desired solution. Then the mobilized atoms undergo impacts which make them enter into combinations among themselves or with other atoms at rest which they struck against in their course. Again I beg pardon, my comparison is very rough, but I scarcely know how otherwise to make my thought understood.

However it may be, the only combinations that have a chance of forming are those where at least one of the elements is one of those atoms freely chosen by our will. Now, it is evidently among these that is found what I call the *good combination*. Perhaps this is a way of lessening the paradoxical in the original hypothesis.

Another observation. It never happens that the unconscious work gives us the result of a somewhat long calculation *all made*, where we have only to apply fixed rules. We might think the wholly automatic subliminal self particularly apt for this sort of work, which is in a way exclusively mechanical. It seems that thinking in the evening upon the factors of a multiplication we might hope to find the product ready made upon our awakening, or again that an algebraic calculation, for example a verification, would be made unconsciously. Nothing of the sort, as observation proves. All one may hope from these inspirations, fruits of unconscious work, is a point of departure for such calculations. As for the calculations themselves, they must be made in the second period of conscious work, that which follows the inspiration, that in which one verifies the results of this inspiration and deduces their consequences. The rules of these calculations are strict and complicated. They require discipline, attention, will, and therefore consciousness. In the subliminal

> self, on the contrary, reigns what I should call liberty, if we might give this name to the simple absence of discipline and to the disorder born of chance. Only, this disorder itself permits unexpected combinations (pp. 392–394).

In addition to these pioneers there of course have been numerous others who in some manner have made a substitute trial and error in a modeled or mnemonic environment an important aspect of their description of thinking. In rough chronology these include Baldwin (1906), Pillsbury (1910), Rignano (1923), Woodworth (1921), Woodworth & Schlosberg (1954), Thurstone (1924), Tolman (1926), Hull (1930), Muenzinger (1938), Miller and Dollard (1941), Craik (1943), Boring (1950), Humphrey (1951), Mowrer (1954), Sluckin (1954), and many others.

OBJECTIONS TO THE MODEL

The Gestalt Protest

The trial-and-error theme in learning was of course one part of the syndrome of ideas against which Gestalt psychology eloquently protested. In spite of this, there is judged to be no inherent conflict between the perspectives of this paper and the Gestalt position. To make this interpretation, it is necessary to regard neither traditional associationism nor Gestalt psychology as discrete integrated wholes, but instead to regard each as congeries of which the parts may be separately accepted or rejected.

The Gestaltists are judged to have validly rejected Thorndike's (1898) description of animal problem solving as solely a matter of overt locomotor trial and error. As this writer (Campbell, 1956b) has argued previously, recognizing vision as a substitute trial-and-error process leads to the expectation that some locomotor problems will be solved by this means, obviating overt trial and error. Even more so does the model for thought. To the present writer the Gestaltists were correct descriptively even though epistemologically the trial-and-error process remains fundamental to discovery. In Wertheimer's (1959) specific contrasts between insightful problem solving and blind trial and error, it is a trial and error of overt manipulation which is involved. Furthermore, as Humphrey (1951) and Woodworth and Schlosberg (1954) point out, the Gestalt descriptions of problem solving provide ample evidence of both fortuitous solutions and misleading "insights." The recurrent Gestaltist protest that even the errorful trials are "intelligent" and that the subsequent trials make use of what was learned through the error are taken here as equivalent to the statements that the problem solver had some valid general knowledge to begin with, and that before acting he employed the substitute trial and error of thought or vision.

The blind-variation-and-selective-retention model of thought joins the Gestaltists in protest against the picture of the learning organisms as a passive induction machine accumulating contingencies. Instead, an active generation and checking of thought-trials, hypotheses, or molar responses is envisaged. The model at the level of thought places essential importance upon internalized selective criteria against which the thought trials are checked. Poincaré's (1913) esthetic criteria and the Gestalt qualities of wholeness, symmetry, organized structure, and the like can be regarded as built-in selective criteria completely compatible with the model. Pringle (1951) for example, has proposed a selective-retention model of central nervous system action in which systematic temporal patterns provide the selective criteria in a resonance process.

Nor does the model here presented specify the nature of the thought trials employed. There must often be a trial and error among possible general principles, or among rational abstractions, or field reorganizations, or recenterings. Both the blind variation and selective survival model, and Gestalt theory emphasize the advantage of breaking out of old ruts, and the disadvantages of set and rote drill (Boring, 1950; Dunker, 1945; Katona, 1940; Luchins, 1942; Wertheimer, 1959). Furthermore, the encountering in thought of an idea which fits can be accompanied at the phenomenal level by a joyful "ahaerlebnis" or a Gestalt experience of "closure," and at the overt performance level by a sudden and stable improvement signifying "insight." There is no essential disagreement here. Nor is the trial-and-error model without phenomenological support. Note the highly similar testimony from the disparate historical citations provided above, especially in the imagery of multitudinous, loosened, agitated, teeming, colliding, and interlocking ideas.

This is not to say that a Gestalt psychologist would be happy with the blind-variation-and-selective-retention description of thought processes. Nor are all aspects of the Gestalt syndrome here accepted. While "insight" is accepted as a phenomenal counterpart of the successful completion of a perhaps unconscious blind-variation cycle, its status as an explanatory concept is rejected, especially as it connotes "direct" ways of knowing. Furthermore, when publicized as a part of an ideology of creativity, it can reduce creativity through giving students a feeling that they lack an important gift possessed by some others, a feeling which inhibits creative effort and increases dependence upon authority. Polya (1945, 1954) has described such an inhibiting tradition in the teaching of mathematics.

Individual Differences and Genius

Another prevalent orientation antithetical in spirit to the blind-variation-and-selective-retention model may be called the "mystique of the creative genius and the creative act." This is related to our deeply rooted tendency toward causal perception (e.g., Heider, 1944), a tendency to see marvelous achievements rooted in equally marvelous antecedents. It takes the form of the "fallacy of accident" and of "*post hoc ergo propter hoc.*" Let a dozen equally brilliant men each propose differing guesses about the unknown in an area of total ignorance, and let the guess of one man prove correct. From the blind-variation-and-selective-survival model this matching of guess and environment would provide us with new knowledge about the environment but would tell us nothing about the greater genius of the one man—he just happened to be standing where lightning struck. In such a case, however, we would ordinarily be tempted to look for a subtle and special talent on the part of this lucky man. However, for the genuinely unanticipatable creative act, our "awe" and "wonder" should be directed outward, at the external world thus revealed, rather than directed toward the antecedents of the discovery. Just as we do not impute special "foresight" to a successful mutant allele over an unsuccessful one, so in many cases of discovery, we should *not* expect marvelous consequents to have had equally marvelous antecedents. Similarly, in comparing the problem-solving efforts of any one person; from the selective survival model it will be futile, in the instance of a genuinely innovative achievement, to look for special antecedent conditions not obtaining for blind-alley efforts: just insofar as there has been a genuine gain in knowledge, the difference between a hit and a miss lies in the selective conditions thus newly encountered, not in talent differences in the generation of the trials.

This is not to deny individual differences in creative intellect. Indeed, the blind-variation-and-selective-retention model of creative thought predicts such talent differences along all of the parameters of the process. This is to emphasize, however, that explanations in terms of special antecedents will very often be irrelevant, and that the causal-interpretative biases of our minds make us prone to such over-interpretations, to *post-hoc-ergo-proper-hoc* interpretations, deifying the creative genius to whom we impute a capacity for direct insight instead of mental flounderings and blind-alley entrances of the kind we are aware typify our own thought processes. Ernst Mach (1896) notes our nostalgia for the directly-knowing genius: "To our humiliation we learn that even the greatest men are born more for life than for science in the extent to which even they are indebted to accident" (p. 175).

What are the ways in which thinkers might be expected to differ, according to the trial-and-error model? First, they may differ in the accuracy and detail of their representations of the external world, of possible locomotions in it or manipulations of its elements, and of the selective criteria. Differences in this accuracy of representation correspond to differences in degree of information and intelligence. Second, thinkers can differ in the number and range of variations in thought trials produced. The more numerous and the more varied such trials, the greater the chance of success. Bain has emphasized the role of fanaticism or extreme dedication in producing large volumes of such explorations. Bain, Souriau, Mach, and Poincaré have all emphasized the role of advance preparation in assembling the elements whose blind permutation and combination make possible a wide range of trials. Many observers have emphasized the role of set and familiarity in reducing the range of variations, and have recommended ways of reducing trial-to-trial stereotypy, as by abandoning the problem for awhile, going on to other things. Devices abound which are designed to increase the likelihood that all permutations be considered and are used by most of us, as in going through the alphabet in finding rhymes or puzzle words. There are no doubt age differences in the rapidity and uninhibited range of thought-trial production. The sociology of knowledge makes an important contribution here: persons who have been uprooted from traditional cultures, or who have been thoroughly exposed to two or more cultures, seem to have the advantage in the range of hypotheses they are apt to consider, and through this means, in the frequency of creative innovation. Thorstein Veblen (1919) has espoused such a theory in his essay on the intellectual preeminence of Jews, as has Robert Park (1928) in writing of the role of "the marginal man" in cultural innovation. (See also Seeman, 1956.) And more generally, it is the principle of variation which leads us to expect among innovators those of personal eccentricity and bizarre behavior. We can also see in this principle the value of those laboratories whose social atmospheres allow wide ranging exploration with great tolerance for blind alley entrances.

The value of wide ranging variation in thought trials is of course vitiated if there is not the precise application of a selective criterion which weeds out the overwhelming bulk of inadequate trials. This editing talent undoubtedly differs widely from person to person, as Poincaré (1908, 1913) has emphasized. With regard to selection criteria, one further point should be made. Much of creative thought is opportunistic in the sense of having a wide number of selective criteria available at all times, against which the thought trials are judged. The more creative thinker may be able to keep in mind more such criteria, and therefore increase his likelihood of achieving a serendipitous (Cannon, 1945; Merton, 1949) advance on a problem tangential to his initial main line of endeavor (e.g., Barber & Fox, 1958). Further areas of individual differences lie in the competence of the retention, cumulation, and transmisison of the encountered solutions.

It need not be expected that these dimensions of talent all go together. In organic evolution, the variation process of mutation and the preservation of gains through genetic rigidity are at odds, with an increase in either being at the expense of the other, and with some degree of compromise being optimum. Just so we might expect that a very pure measure of innovative range in thought and a very pure measure of rote memory might be even negatively correlated, as Saugstad (1952) seems to have found, and similarly for innovative range and selective precision. Such considerations suggest complementary combinations of talent in creative teams, although the uninhibited idea-man and the compulsive edit-and-record type are notoriously incompatible office mates.

Notice regarding the individual differences thus described that while they do make creative innovation much more likely on the part of some individuals than others, they do not place the joys of creative innovation beyond the reach of the less gifted. Indeed, looking at large populations of thinkers, the principles make it likely that many important contributions will come from the relatively untalented, undiligent, and uneducated, even though on an average contribution per capita basis, they will contribute much less, points which Souriau (1881) has noted. The intricacy of the tradition to which innovation is being added of course places limitations in this regard.

The Enormous Domain of Possible Thought-Trials to be Searched

A final type of objection to the blind-variation-and-selective-retention model of thought needs to be considered. This objection is to the effect that the domain of possible thought trials is so large that the solution of a given problem would take an impossibly long time were a search of all possibilities to be involved, either through a systematic scanning of all possibilities where these are enumerable, or through a random sampling of the universe of possibilities. Time and trial estimates thus based can be overwhelming, as Kurt Lasswitz's story "The Universal Library" (1958) dramatically illustrates. Other parodies of our model occur in literature as far back as Swift's portrait of the Academy of Lagado in *Gulliver's Travels* (1941, pp. 166–169). (Ley, 1958, traces such ideas back to Lully, ca. 1200.) Newell, Shaw, and Simon (1958a, 1958b) refer in this vein to what they call the "British Museum Algorithm," i.e., the possibility of a group of trained chimpanzees typing at random producing by chance in the course of a million years all of the books in the British Museum. Such parodies seem effectively to reject the blind-variation-and-selective-retention model through a *reductio ad absurdum*. Needless to say, such a rejection is not accepted in the present paper. As a matter of fact, it is judged to be in the same class as parallel objections to the theory of natural selection in evolution. Similar features in these two instances make the accidentalist interpretation more acceptable.

1. Neither in organic evolution nor in thought are all problems solved, nor all possible excellent solutions achieved. There is no guarantee of omniscience. The knowledge we do encounter is achieved against terrific odds. (Those advocating heuristically-programed problem-solving computers are careful not to guarantee solutions, and this modesty should be extended to all models of creative thought.)

2. The tremendous number of nonproductive thought trials on the part of the total intellectual community must not be underestimated. Think of what a small proportion of thought becomes conscious, and of conscious thought what a small proportion gets uttered, what a still smaller fragment gets published, and what a small proportion of what is published is used by the next intellectual generation. There is a tremendous wastefulness, slowness, and rarity of achievement.

3. In biological evolution and in thought, the number of variations explored is greatly reduced by having *selective criteria imposed at every step.* Thus mutant variations on nonadaptive variations of the previous generation are never tested—even though many wonderful combinations may be missed therefore. Some of the "heuristics" currently employed in logic and chess playing machines (Newell, Shaw, & Simon, 1958a, 1958b) have the similar effect of evaluating all next-possible moves in terms of immediate criteria, and then of exploring further variations upon only those stems passing the screening of each prior stage. It is this strategy of cumulating selected outcomes from a blind variation, and then exploring further blind variations only for this highly select stem, that, as R. A. Fisher has pointed out (e.g., 1954, p. 91) makes the improbable inevitable in organic evolution. This strategy is unavoidable for organic evolution, but can obviously be relaxed in thought processes and in machine problem solving. However, the Pandora's box of permutations opened up by such relaxation can be used to infer that, in general, thought trials are selected or rejected within one or two removes of the established base from which they start. In constructing our "universal library" we stop work on any volume as soon as it is clear that it is gibberish.

4. When we make estimates of the number of permutations which would have to be culled to obtain a given outcome, we often assume that problem solving was undertaken with that one fixed goal in mind. This overlooks the opportunistic, serendipitous course of organic evolution and of much of creative thinking. The likelihood of a productive thought increases with the wider variety of reasons one has for judging a given outcome "interesting." To neglect this opportunistic multipurposedness gives one a poor base for estimating the probability of encountering the one outcome hit upon and recorded. Thus when Newell, Shaw, and Simon's "Logic Theorist" (1958a, 1958b) sets out to prove the 60-odd theorems in a given chapter of *Principia Mathematica,* it may face a more formidable task than did Whitehead and Russell in generating them, if, except for the dozen classic theorems reproduced, Whitehead and Russell were otherwise free to record every deduction they encountered which seemed "interesting"

or "nontrivial." Wigglesworth (1955) has noted this strategy on the part of "pure" scientists, in commenting on the relationship between pure and applied scientists in wartime:

> In the pure science to which they were accustomed, if they were unable to solve problem A they could turn to problem B, and while studying this with perhaps small prospect of success they might suddenly come across a clue to the solution of problem C (p. 34).

In presenting their case for adding "heuristics" to the program of the "Logic Theorist," Newell, Shaw, and Simon have emphasized the inadequacy of blind trial and error. So has Miller (1959) in advocating the heuristic of searching backward from the goal.[4] There is, however, no essential disagreement between their point of view and the one offered here. By adding heuristics mechanical thought processes have indeed been made more like those of human beings, both in adequacy and type of errors. Such innovations have obviated the protests of those such as Wisdom (1952) and Mays (1956) who, while conceding that machines could choose good moves at chess or solve logic problems, have found the machines failing to imitate life just in their orderly inspection of all possibilities. Newell, Shaw, and Simon recognize that a machine which would develop its own heuristics would have to do so by a trial and error of heuristic principles, with no guarantee that any would work. They further recognize that possession of an effective heuristic represents already achieved general knowledge about the domain under search, and that adding to this general knowledge will be a blind search process. (The devices of learning and vision and of coding environmental possibilities for thought-search all represent heuristics in this sense.) They might also agree that most heuristic devices will be limited to the specific domain of their discovery, and can only be extended to other domains on a trial basis. They would probably also agree that no problem solving process will be "direct." The disagreements I have with their excellent paper on the processes of creative thinking (Newell, Shaw, & Simon, 1958b) are thus minor matters of emphasis, but may be worth stating nonetheless, to further clarify the position here advocated. They say, for example:

---

[4] Miller (1959, pp. 244–246) is wrong in implying that the strategy of working backward from the goal eliminates the necessity of symbolic trial and error in creative thinking. His mistake comes from assuming that only one path leads into a goal or subgoal. In the spatial locomotion problems from which his concrete illustrations come, and for the logic problems used by Newell, Shaw, and Simon, the paths into any position are not singular, but are instead typically as numerous as paths leading out. A *pure* strategy of working from the goal back to the start would thus involve exploring just as many permutations as the pure strategy of exploring all paths from the start position. However, there is a useful strategy available to symbolic trial and error and not to overt trial and error, in working concurrently from both ends. This produces a dramatic advantage in the number of permutations generated, and a smaller but still substantial gain in the number of comparisons. In the instance suggested by Miller in which each locus branches into 4 alternatives and in which the start and goal turn out to lie 7 stages apart, either of the pure strategies would generate $4 + 4^2 + 4^3 + 4^4 + 4^5 + 4^6 + 4^7$ or 21,844 permutations (neglecting the probable achievement of success before exhausting the 4^7 generation of alternatives). For the mixed approach, the junction would be encountered at the third stage away from each end, or when $2(4 + 4^2 + 4^3)$ or 168 alternatives had been examined. In the pure strategy, there would be 21,845 comparisons involved, that is, the start position and each subsequent alternative would be compared with the goal. In the mixed strategy, many more comparisons per alternative are required. Each permutation must be compared with each of the current and previous permutations on the other stem, which in this instance amounts to 7,225 comparisons. If the comparisons are regarded as equally costly as the generating and storing of alternatives, the savings of the mixed approach over either of the pure approaches would amount to approximately 1 to 6, a very handsome gain for any heuristic. This gain is larger as the number of branches at each stage increases. Advocacy of the heuristic of working backwards in what is essentially this mixed form is present in Polya (1945), Wisdom (1952), and Newell, Shaw, and Simon (1958a, 1958b). In none of these is it claimed that trial and error is eliminated, while all point to the reduction in trials which it can achieve.

> We have given enough estimates of the sizes of the spaces involved . . . to cast suspicion upon a theory of creativity which places its emphasis upon increase in trial and error (p. 63).

The blind-variation-and-selective-retention model unequivocally implies that *ceteris paribus,* the greater the heterogeneity and volume of trials the greater the chance of a productive innovation. Doubling the number of efforts very nearly doubles the chance of a hit, particularly when trials are a small part of the total domain and the repetitiousness among trials is low. But they too recognize unconventionality and no doubt numerosity as a necessary, if not a sufficient condition of creativity (1958b, p. 62). What they would validly stress, is the very frequent tactical advantage of a trial and error of general strategic principles over a trial and error involving no classificatory effort nor attempt to use clues, and, once such general heuristics have been discovered, the advantages of a hierarchized trial and error process. The advantage of such a strategy depends upon the ecology, of course, but we are in general justified in expecting solutions to be nonrandomly distributed, and to show significant contingencies with prior clues. Polya (1945, 1954) has, of course, been a major source of inspiration for all efforts to introduce heuristics into problem solving, and for him a trial and error approach is a heuristic of fundamental importance.

Another minor point of disagreement may be mentioned. In their efforts to consider how a "Logic Theorist" might be programed to learn a general heuristic from hindsight they propose that it keep a record of the outcomes of all past trials, successful and unsuccessful, in order to be able to scan its experience for general principles of strategy (1958b). Implementing this would put a tremendous strain upon memory storage, and would introduce a scanning process as time consuming as the original search process which produced the record. The strategy of organic evolution is to keep a record only of what works, even at the expense of repeating its errors. The general preponderance of wrong tries at every level, plus problems of memory glut and access, suggests a similar strategy for all knowledge processes. Heuristics can probably best be learned through a trial and error of heuristics, tried on new problem sets rather than old.

STATUS AS A THEORY

At the level here developed, one might better speak of an "orientation to," or a "perspective on" creative thought processes, rather than a "theory of." At many points, this perspective merely points to problems, rather than taking that step toward theory of providing guesses at answers: e.g., for the processes here outlined to be possible, theory must not only provide several memory processes, but most importantly, must specify a possible mechanism for the trial-and-error search of these. From such specifications will come the subtlety of prediction characteristic of a developed theory.[5] While the perspective even in its very general state has some unequivocal empirical implications, the major advantage to it may be metaphysical, or at least metatheoretical. Like the theory of natural selection in organic evolution, it provides an understanding of marvelously purposive processes without the introduction of teleological metaphysics or of pseudocausal processes working backward in time.

Note that there are still ambiguities about the status *as theory* of the well established principle of organic evolution through natural selection, even though now buttressed with the detailed genetic model of the variation and retention processes. Scriven (1959) has called it an explanatory rather than a predictive theory. While the present writer feels that Scriven has somewhat undervalued the experimental studies of evolution with viruses, bacteria, and insects, and may have mistaken the past absence of experimental settings appropriate to testing the theory for an inherent attribute of the type of theory per se; he has called attention, nonetheless, to some serious problems. Even Sewall Wright (1960), whose statistical genetic theory of evolution has added subtle details to the overall description of the process, has commented in a similar vein:

> The theory is deterministic only in an exceedingly limited sense. It is essentially a theory of the conditions favorable for an ever continuing process that is essentially unpredictable in its details. There can be no formula for serendipity (p. 148).

The problems which a selective retention theory of creative thinking shares with that of evolution include the following:

1. The basic insight, so useful and so thrilling when first encountered, is close to being an analytic tautology rather than a synthetic description of process: if indeed variations occur which are differentially selected and propagated, then an evolutionary process toward better fit to any set of consistent selective criteria is inevitable.
2. For most applications of the selective retention model, the variation is taken as a descriptive given, as an unexplained part of the explanation. While other predictive theories likewise depend upon unexplained processes at a more molecular level, this instance may seem evasive at a particularly crucial point, and has indeed been taken as a denial of determinism or as a rival metaphysic of "spontaneous change" as presumptive as a teleological one. We are currently getting detailed deterministic explanations of the mutation of genes, but until something comparable is available to predict the generation of heterogeneous thought trials, this constitutes a weakness of the model.
3. The biological study of the evolution of any species takes the form of a post hoc reconstruction of a unique, "undetermined," historical process. The achievement of any general regularities must be probabilistic in the extreme. Studies starting from specific spectacular achievements in creative thought must be similar in nature.
4. The theory suffers from the multiplicity

---

[5] This paper does not attempt to review theories in this area. Citations to the important contributions of Pringle (1951) and Hebb (1949), do not begin to represent this literature. Note the special problem of a brain analogue to switching. Ashby's (1952) model and most computer memory search involves a spatial displacement of solids impossible in the brain, as in the stepping switch or the rotation of a magnetic memory drum. Computer memory search processes making use of timed pulses require a precision of timing dependent upon a stability of dimension presumably not available in the brain and usually if not always dependent upon a clock within which actual spatial displacement of solids takes place. It is for these reasons that Pringle's (1951) theory seems particularly promising, and one wonders why it has not been more used, or if not usable, more publicly refuted. See Pribram (1959) for a recent contribution to the problem of appropriate brain process models.

of possible mediations it posits. Where there are gaps in the historical knowledge, the theory makes available an embarrassing surfeit of possible reconstructions. In this, and in contrast to the successful theories of macrophysics, the theory is less self-disciplining, less specific in its predictions, more evasive of potential disconfirmation. This is perhaps the most important of Scriven's (1959) points, and one equally applicable to the theory of creative thinking.

5. In the usual applications, the environment is not described or describable prior to the organismic achievement of adaptation to that environment. Whereas the theory deals with an iterative process whereby an organism adapts to (achieves knowledge of) an independent environment, the evidences as to the organismic form and environmental parameters are often confounded, in that the same data series is used to infer both. While this is not so for laboratory studies of trial-and-error learning, it is particularly apt to be so for any study of truly great creative thinking in science. (See Campbell, 1959, p. 157 for epistemological citations to this problem.)

There is in addition, a serious problem which the blind-variation-and-selective-retention theory of creative thought faces which is not present in comparable degree in the modern theory of organic evolution. This is the unfavorable ratio of hypothesized unobservable processes to observable input-output variables. Note that even in its sketchy form here given, some 6 to 14 or more separately variable parameters are implied. These include: (*a*) A mnemonic representation of environment, varying perhaps in scope, accuracy, and fineness of detail; (*b*) A mnemonic search or thought-trial process, varying in the accuracy with which it represents potential overt exploration; (*c*) A thought-trial generating and changing process, varying in rate, heterogeneity, idiosyncrasy, and lack of repetitiousness among successive thought trials; (*d*) Selective criteria, varying in their number, accuracy of representation of environmental contingencies, and precision, sharpness, or selection ratio; (*e*) A preservation or propagation process, providing a retention for selected thought trials of a quite different order from the memory traces of the nonselected ones, varying perhaps in accuracy and accessibility; (*f*) A reality testing process in which the selected thought trials are checked out by overt locomotion in the external environment, varying perhaps in sensitivity to disconfirming feed-back.

This inventory of weaknesses does not, of course, represent argument in favor of rival theories of creative thought, which are judged to be still more amorphous, still less adequate. And as Duncan (1959) makes clear in reviewing the research literature on human problem solving, for all theories there is lacking a disciplined relation both to experimental undertakings and to findings. Even in its present form, however, the theory contains many empirical implications. Manipulation of any one of the 14 variables just listed should increase the number of creative products, providing the other variables can be held constant. Predictions of this order have been specified in the discussion of individual differences. Particularly characteristic are the unequivocal predictions regarding the volume and heterogeneity of thought trials. *Ceteris paribus,* a creative solution is more likely the longer a problem is worked upon, the more variable the thought trials, the more people working on the problem independently, the more heterogeneous these people, the less the time pressure, etc.

SUMMARY

This paper has attempted to make the psychological and epistemological point that all processes leading to expansions of knowledge involve a blind-variation-and-selective-retention process. Processes, such as vision and

thought, substituting for an overt trial and error are of couse acknowledged. But each of these are interpreted as containing in its very workability wisdom about the environment obtained originally by the blind variation of mutation and natural selection. In addition, each contains a blind-variation-and-selective-retention process at its own level. Supporting the effort to interpret all knowledge processes in this light has been an emphasis upon the tremendous gain in knowledge in the course of evolution and history, a gain which can only be explained by a continual breakout from the bounds of what was already known, a breakout for which blind variation provides the only mechanism available.

The application of this perspective to the process of creative thought antedates its application to trial-and-error learning in animals and to organic evolution, and is illustrated through quotations from Bain, Souriau, Mach, and Poincaré. The model is not in disagreement with the bulk of the Gestalt description of problem solving, nor the work on heuristically programed problem-solving computers. However, there is an effort to root out a prevailing implicit belief in the possibility of "direct" or "insightful" creative thought processes.

REFERENCES

ASHBY, W. R. *Design for a brain.* New York: Wiley, 1952.

BAIN, A. *The senses and the intellect.* (3rd ed.) New York: Appleton, 1874.

BALDWIN, J. M. *Mental development in the child and the race.* New York: Macmillan, 1900.

BALDWIN, J. M. *Thought and things, or genetic logic.* Vols. I-III. New York: Macmillan, 1906–1911.

BARBER, B., & FOX, R. C. The case of the floppy-eared rabbits: An instance of serendipity gained and serendipity lost. *Amer. J, Sociol.*, 1958, **64**, 128–136.

BORING, E. G. Great men and scientific progress. *Proc. Amer. Phil. Soc.*, 1950, **94**, 339–351.

CAMPBELL, D. T. Adaptive behavior from random response. *Behav. Sci.*, 1956, **1**, 105–110. (a)

CAMPBELL, D. T. Perception as substitute trial and error. *Psychol. Rev.*, 1956, **63**, 330–342. (b)

CAMPBELL, D. T. Methodological suggestions from a comparative psychology of knowledge processes. *Inquiry,* 1959, **2**, 152–182.

CANNON, W. B. *The way of an investigator.* New York: Norton, 1945.

CRAIK, K. J. W. *The nature of explanation.* Cambridge: Cambridge Univer. Press, 1943.

DEUTSCH, J. A., & CLARKSON, J. K. The nature of the vibrato and the control loop in singing. *Nature,* 1959, **183**, 167–168.

DITCHBURN, R. W. Eye-movement in relation to retinal action. *Optica Acta,* 1955, **1**, 171–176.

DUNCAN, C. P. Recent research on human problem solving. *Psychol. Bull.*, 1959, **56**, 397–429.

DUNKER, K. On problem-solving. *Psychol. Monogr.*, 1945, **58**(5, Whole No. 270).

FISHER, R. A. Retrospect of the criticism of the theory of natural selection. In J. Huxley, A. C. Hardy, & E. B. Ford (Eds.), *Evolution as a process.* London: Allen & Unwin, 1954.

GRIFFIN, D. R. *Listening in the dark.* New Haven: Yale Univer. Press, 1958.

HADAMARD, J. *The psychology of invention in the mathematical field.* Princeton: Princeton Univer. Press, 1945.

HEBB, D. O. *The organization of behavior.* New York: Wiley, 1949.

HEIDER, F. Social perception and phenomenal causality. *Psychol. Rev.*, 1944, **51**, 358–374.

HULL, C. L. Knowledge and purpose as habit mechanisms. *Psychol. Rev.*, 1930, **37**, 511–525.

HUMPHREY, G. *Thinking.* London: Methuen, 1951.

JENNINGS, H. S. *The behavior of the lower organisms.* New York: Columbia Univer. Press, 1906.

KATONA, G. *Organizing and memorizing.* New York: Columbia Univer. Press, 1940.

KELLOGG, W. N. Echo-ranging in the porpoise. *Science*, 1958, **128**, 982–988.

LASSWITZ, K. The universal library. (First published 1901) In C. Fadiman (Ed.), *Fantasia mathematica.* New York: Simon & Schuster, 1958.

Ley, W. Postscript to "The universal library." In C. Fadiman (Ed.), *Fantasia mathematica.* New York: Simon & Schuster, 1958. Pp. 244–247.

Luchins, A. Mechanization in problem solving: The effect of Einstellung. *Psychol. Monogr.,* 1942, **54**(6, Whole No. 248).

Mach, E. On the part played by accident in invention and discovery. *Monist,* 1896, **6**, 161–175.

Mays, W. Cybernetic models and thought processes. *Proceedings of the First International Congress on Cybernetics, Namur.* Paris: Gauthier-Villard, 1956.

Merton, R. K. *Social theory and social structure.* Glencoe, Ill.: Free Press, 1949.

Miller, N. E. Liberalization of basic S-R concepts: Extensions to conflict behavior, motivation, and social learning. In S. Koch (Ed.), *Psychology: A study of a science.* Vol. 2. *General systematic formulations, learning, and special processes.* New York: McGraw-Hill, 1959.

Miller, N. E., & Dollard, J. *Social learning and imitation.* New Haven: Yale Univer. Press, 1941.

Mowrer, O. H. Ego psychology, cybernetics, and learning theory. In *Kentucky symposium: Learning theory, personality theory, and clinical research.* New York: Wiley, 1954. Pp. 81–90.

Muenzinger, K. F. Vicarious trial and error at a point of choice: I. A general survey of its relation to learning efficiency. *J. genet. Psychol.,* 1938, **53**, 75–86.

Newell, A., Shaw, J. C., & Simon, H. A. Elements of a theory of human problem solving. *Psychol. Rev.,* 1958, **65**, 151–166. (a)

Newell, A., Shaw, J. C., & Simon, H. A. *The process of creative thinking.* University of Colorado 1958 symposium on cognition. *Rand Corp. Rep.,* 1958, No. P-1320. (b)

Park, R. E. Human migration and the marginal man. *Amer. J. Sociol.,* 1928, **33**, 881–893.

Pillsbury, W. B. *The psychology of reasoning.* New York: Appleton, 1910.

Platt, J. Amplification aspects of biological response and mental activity. *Amer. Scient.,* 1956, **44**, 181–197.

Platt, J. Functional geometry and the determination of pattern in mosaic receptors. In H. P. Yockey, R. L. Platzman, & H. Quastler (Eds.), *Symposium on information theory in biology.* New York: Pergamon, 1958. Pp. 371–398.

Poincaré, H. L'invention mathematique. *Bull. Inst. Gen. Psychol.,* 1908, **8**, 175–187.

Poincaré, H. Mathematical creation. In H. Poincaré, *The foundations of science.* New York: Science Press, 1913.

Polya, G. *How to solve it.* Princeton, N. J.: Princeton Univer. Press, 1945.

Polya, G. *Mathematics and plausible reasoning.* Vol. I. *Induction and analogy in mathematics.* Vol. II. *Patterns of plausible inference.* Princeton, N. J.: Princeton Univer. Press, 1954.

Pribram, K. H. On the neurology of thinking. *Behav. Sci.,* 1959, **4**, 265–287.

Pringle, J. W. S. On the parallel between learning and evolution. *Behaviour,* 1951, **3**, 175–215.

Pumphrey, R. J. Hearing. In *Symposia of the Society for Experimental Biology.* IV. *Physiological mechanisms in animal behavior.* New York: Academic Press, 1950.

Riggs, L. A., Armington, J. C., & Ratliff, F. Motions of the retinal image during fixation. *J. Opt. Soc. Amer.,* 1954, **44**, 315–321.

Rignano, E. *Psychologie du raisonnement.* Paris: Alcan, 1920. (Reprinted: *The psychology of reasoning.* New York: Harcourt, Brace, 1923).

Saugstad, P. Incidental memory and problem-solving. *Psychol. Rev.,* 1952, **59**, 221–226.

Scriven, M. Explanation and prediction in evolutionary theory. *Science,* 1959, **130**, 477–482.

Seeman, M. Intellectual perspective and adjustment to minority status. *Soc. Probl.,* 1956, **3**, 142–153.

Simon, H. A. *Models of man.* New York: Wiley, 1957.

Sluckin, W. *Minds and machines.* Harmondsworth: Penguin, 1954.

Souriau, P. *Theorie de l'invention.* Paris: Hachette, 1881.

Swift, J. *Gulliver's travels.* (Originally published 1726) Oxford: Blackwell, 1941.

Thorndike, E. L. Animal intelligence: An experimental study of the associative processes in animals. *Psychol. Rev. Monogr. Suppl.,* 1898, **2**(4, Whole No. 8).

Thurstone, L. L. *The nature of intelligence.* New York: Harcourt, Brace, 1924.

Tolman, E. C. A behavioristic theory of ideas. *Psychol. Rev.,* 1926, **33**, 352–369.

VEBLEN, T. The intellectual preeminence of Jews in modern Europe. *Pol. sci. Quart.* 1919, **34**, 33–42.

WERTHEIMER, M. *Productive thinking.* New York: Harper, 1959.

WIGGLESWORTH, V. B. The contribution of pure science to applied biology. *Ann. appl. Biol.,* 1955, **42**, 34–44.

WISDOM, J. O. Symposium: Mentality in machines. In Aristotelian Society, Suppl. Vol. 26, *Men and machines.* London: Harrison, 1952. Pp. 1–26.

WOODWORTH, R. S. *Psychology.* New York: Holt, 1921.

WOODWORTH, R. S., & SCHLOSBERG, H. *Experimental psychology.* (Rev. ed.) New York: Holt, 1954.

WRIGHT, S., Physiological genetics, ecology of populations, and natural selection. *Perspectives Biol. Med.,* 1959, **3**, 107–151. (Reprinted: Tax, S. (Ed.) *Evolution after Darwin.* Vol. I. *The evolution of life.* Chicago: Univer. Chicago Press, 1960.)

(Received October 21, 1959)

CONTEMPORARY PSYCHOLOGY

. . . APA's journal of book reviews

- Critical reviews of the most recent books in psychology and related areas
- The Editor's page: comments on news in the publishing world, on criticism, on literary style
- A letters-to-the-editor section
- Instructional Media: reviews and comments on films, TV programs, workbooks, teaching machines

AMERICAN PSYCHOLOGICAL ASSOCIATION

1333 Sixteenth St., N.W.
Washington 6, D. C.

JOURNAL OF EDUCATIONAL PSYCHOLOGY

This journal is now published by the American Psychological Association. In 1958 it became a bimonthly with issues in February, April, June, August, October, and December. Contents include articles on problems of teaching, learning, and the measurement of psychological development.

All back issues and subscriptions up to and including the May 1957 issue are the property of Warwick and York, Inc., 10 East Centre Street, Baltimore 2, Maryland.

1961 Subscription $10.00
(Foreign, $10.50)

Single Copies, $2.00

Direct new subscriptions and renewals to:

AMERICAN PSYCHOLOGICAL ASSOCIATION
Publications Office
1333 Sixteenth Street, N. W.
Washington 6, D. C.